ERDC/CERL TR-00-20

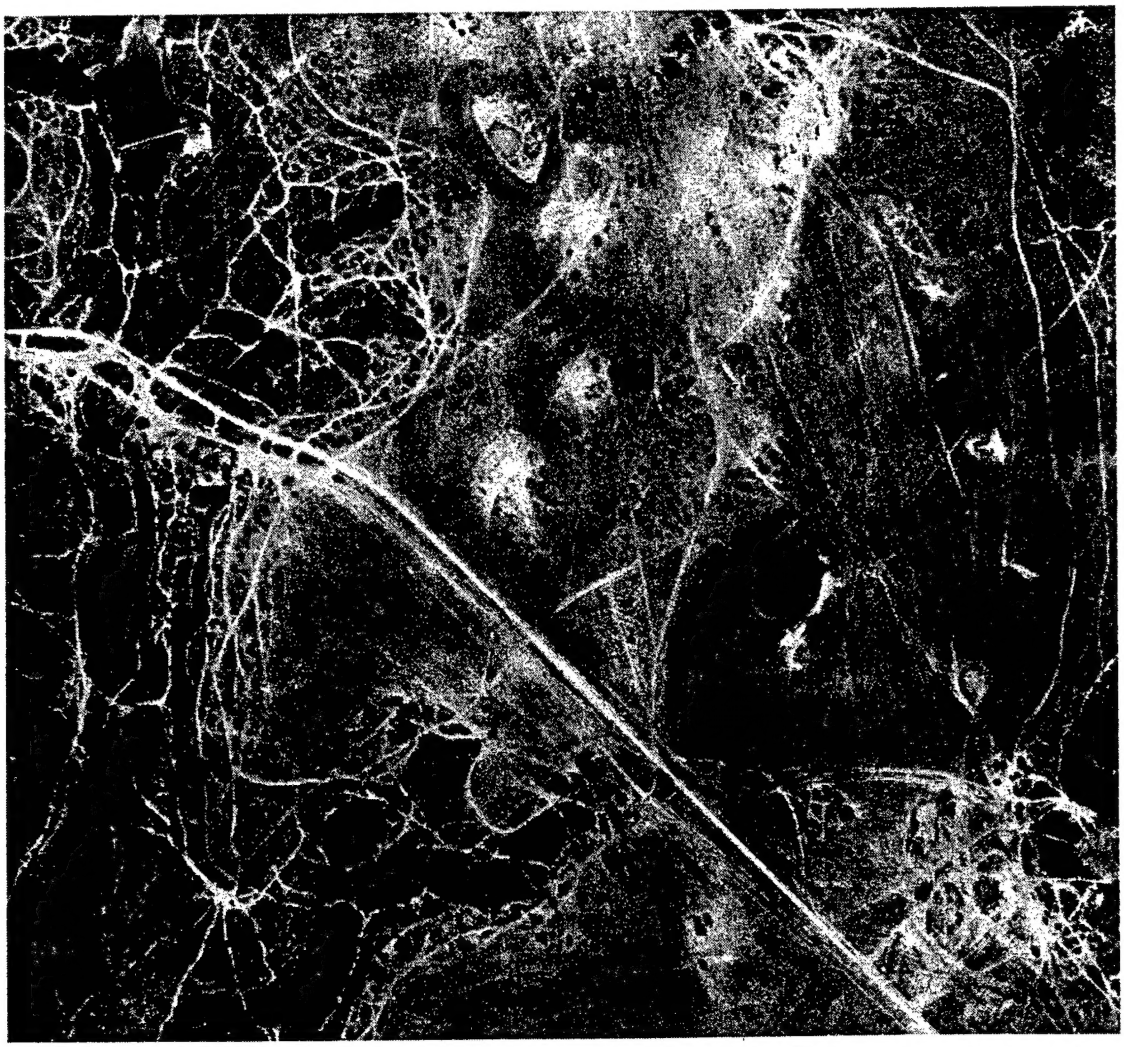
Construction Engineering Research Laboratory

US Army Corps of Engineers®
Engineer Research and Development Center

Quality Assurance/Quality Control Procedures for ITAM GIS Databases

by Douglas M. Johnston, Diane M. Timlin, Diane L. Szafoni, Jason J. Casanova, and Kelly M. Dilks

August 2000



Foreword

This study was conducted for the Assistant Chief of Staff for Installation Management, Directorate of Environmental Programs under 62720A917, work unit number BF8, "QA/QC Procedures for Fort Hood ITAM Data." The technical reviewer was Gordon Weith, Army Training Support Center.

The work was performed by the Land and Heritage Conservation Branch (CN-C) of the Installations Division (CN), Construction Engineering Research Laboratory (CERL). The CERL Principal Investigator was Kelly M. Dilks. Part of this work was done by the Geographic Modeling Systems Laboratory and the University of Illinois at Urbana/Champaign (UIUC) under contract No. DACA88-97-D-0004. The technical editor was Linda L. Wheatley, Information Technology Laboratory. Robert E. Riggins is Chief, CN-C, and Dr. John T. Bandy is Chief, CN. The associated Technical Director is Dr. William D. Severinghaus. The Acting Director of CERL is Dr. Alan Moore.

CERL is an element of the U.S. Army Engineer Research and Development Center (ERDC), U.S. Army Corps of Engineers. The Director of ERDC is Dr. James R. Houston and the Commander is COL James S. Weller.

Contents

List of Figures and Tables

Figures

Tables

1 Introduction

Background

Geographic information creates unique challenges and opportunities for realizing the mission objectives of the U.S. Army, both in tactical operations and in readiness preparation. The Army relies heavily on regular and accurate surveys of cultural, biological, and geological characteristics of its installations to maintain and improve its capabilities. It uses geographic information to document the locations and characteristics of infrastructure such as roads, utilities, and ranges. Geographic information provides the basis for management of training areas on military installations including soil and vegetative conditions critical to training safety and realism, environmental monitoring and compliance, maintenance and management investments, and real property management.

It is estimated the Federal Government spends over $3 billion on spatial data in each fiscal year and Geographic Information Systems (GIS) have been implemented at nearly every U.S. Army installation in the United States. Since 1985, the Army has invested millions of dollars in this implementation, including: hardware procurement, GIS software development, GIS training, and GIS data development.

In spite of the large investment in digital geographic information and systems, numerous challenges to effective creation, analysis, and delivery of geographic information exist. For example, rapidly growing sources of geospatial data from high-resolution satellite or airborne sensors, global positioning satellite-based data, and a multitude of derivative geospatial products produced by a variety of government and private sources pose severe challenges for integrating data. Difficulties that arise include duplicate information and contradictory information (differences in geographic and attribute descriptions of the same features).

Other challenges are introduced by the technology. Issues include interoperability between data representations, computer hardware and software systems, and networked or distributed data and processing. Another set of challenges focuses on the management of the data. Current techniques for storing and managing data are not designed to handle diverse sources of information,

differences in quality, tracking of changes, and other needs introduced by the rapid expansion of development and use of geographic information.

By its very nature, the use of data to represent features and processes of the landscape implies approximation, or error, in representation. Thus, while improving the accuracy of sensors and other data collection methods is important, the thrust of the challenges identified above can be summarized into one requirement: the need for rigorous documentation of the characteristics and quality of data.

GIS are an integral part of the Army, including the Integrated Training Area Management (ITAM) program. To realize the maximum utility of GIS to Army programs and personnel, this requirement must be addressed.

Objectives

The primary goal of this research was to develop and test methodology to assess, report, and improve the quality of spatial data used in Army installation ITAM databases. The specific tasks included: identification and performance of Quality Assurance and Quality Control (QA/QC) procedures on Fort Hood ITAM GIS data layers; documentation of core ITAM GIS data layers using the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata.

Approach

The approach was a process of assessing the status and quality of selected existing data sets based on current standards and the research literature, investigating methods and resource requirements to improve the quality of these data sets, and documentation of the findings. Results of the assessment and improvements are reported according to the FGDC Content Standard for Digital Geospatial Metadata using the CORPSMET95 software* and other documentation. To initiate this effort and provide a framework for future efforts, the project developed and tested procedures for performing a post-development assessment of the quality of selected data from the ITAM data set at Fort Hood,

* For more information on CORPSMET95, see pp 45-46.

Texas. The methodology for QA/QC presented in Chapter 3, incorporates practical and accepted approaches promoted in various standards (such as the Federal Information Processing Standard [FIPS] Spatial Data Transfer Standard) and in current research literature.

This report is organized in the following manner. Chapter 2 presents an overview of relevant standards and procedures currently in use, as well as a summary of the characteristics or components of quality assessment. Chapter 3 discusses in detail metrics for assessment of data and proposes an assessment method. Chapter 4 describes the application of the assessment method to the ITAM GIS data at Fort Hood, Texas. Chapter 5 presents the results from the assessment, while Chapter 6 provides a summary, conclusions, and recommendations.

Scope

The project uses a combination of theoretical work on the nature of geospatial data quality and a case study to assess the requirements for conducting a data quality assessment. The project develops and tests procedures for performing assessment of the quality of an existing data set (specifically, selected data from the ITAM GIS database in use at Fort Hood, Texas). The procedures are intended to be generalizable to other installation ITAM data sets.

Specific tasks included in the project scope were to:

- Individually document a selected, representative set of the core ITAM GIS data layers available at Fort Hood, utilizing the FGDCs content standard for digital geospatial metadata.
- Develop QA/QC procedures for the ITAM GIS data layers. These procedures include positional accuracy, attribute accuracy, completeness, and consistency.
- Conduct QA/QC procedures on the ITAM GIS data layers including assessment, editing, evaluation, and testing procedures to improve the quality of an existing data set. This task entailed investigating and/or developing post-processing mechanisms for addressing documentation of data quality. Updates to the data set will be accompanied by updates to the quality report.
- Document approach, resource requirements, and results with CORPSMET95.

The final deliverables for the project were to:

- Test methodology for quality assessment
- Run quality assessments for selected data sets
- Document metadata for selected data sets, completed using CORPSMET95

The resulting methods and procedures examine and evaluate the quality of a data set and produce a quality report. The quality report describes the data set, how it was derived, what data it intends to represent, and how well it represents it. Key components are the descriptions and quantitative measures of a data set's quality. The content of the report will provide critical information needed to enable a user to objectively determine the data set's fitness for use.

Mode of Technology Transfer

Mechanisms for technology transfer of this product include delivery of data sets and data reports to Fort Hood ITAM personnel, and conference presentations on the work conducted. Future work will examine the development of software tools to facilitate the conduct of QA/QC procedures and improvements to databases.

2 Overview of Geospatial Data Standards

This chapter is an overview of current concepts pertaining to the description and analysis of geospatial data quality. Numerous organizations, committees, and the academic research community are investigating or have promulgated standards and procedures for assessing and documenting the characteristics of data sets to promote interoperability and useability of geospatial data.

Standards fall into several categories. Among them are data file structure and format (e.g., the Spatial Data Transfer Standard [STDS]), domain and data schema standards (e.g., the Spatial Data Standard [SDS]), or structures for describing data sets (regardless of any other standard used to implement a data set, e.g., the FGDC's Content Standard for Digital Geospatial Metadata [CSDGM]), and accuracy (e.g., National Standard for Spatial Data Accuracy [NSSDA]). These standards provide a baseline of comparison for establishing quality assessment of a data set. In this study, the focus is on defining elements of geospatial data quality, investigating methods for assessment, and ensuring conformance of data through application of existing standards.

Standards Organizations

Federal Geographic Data Committee

The FGDC is an interagency committee, organized in 1990 under OMB Circular A-16 to promote the coordinated use, sharing, and dissemination of geospatial data on a national basis. The FGDC is composed of representatives from Cabinet level and independent Federal agencies. The FGDC has established a standards process that is followed in all of its research areas. Subcommittees focus on specific information types to advance the development of standard data models including bathymetry, cadastral, geologic, transportation, vegetation, and wetlands. Working groups focus on information distribution issues such as clearinghouses, metadata, data formats, and transfer.

National Spatial Data Infrastructure (NSDI)

Executive Order (EO) 12906 of April 1994 called for the establishment of a coordinated NSDI "to support public and private sector applications of geospatial data." The EO described activities that were to be undertaken by the Federal Government to promote data sharing among Federal, state, and local governments, citizens, private sector organizations, and academia. The purpose of these activities was to make accurate and timely geographic data readily available to support sound decisions over a geographic area, to do so with minimum duplication of effort, and at a reasonable cost. The FGDC, composed of 14 agencies that produce and use geographic data, was charged with coordinating the Federal Government's development of the NSDI. One of the major initiatives was the development of standards for data documentation, collection, and exchange.

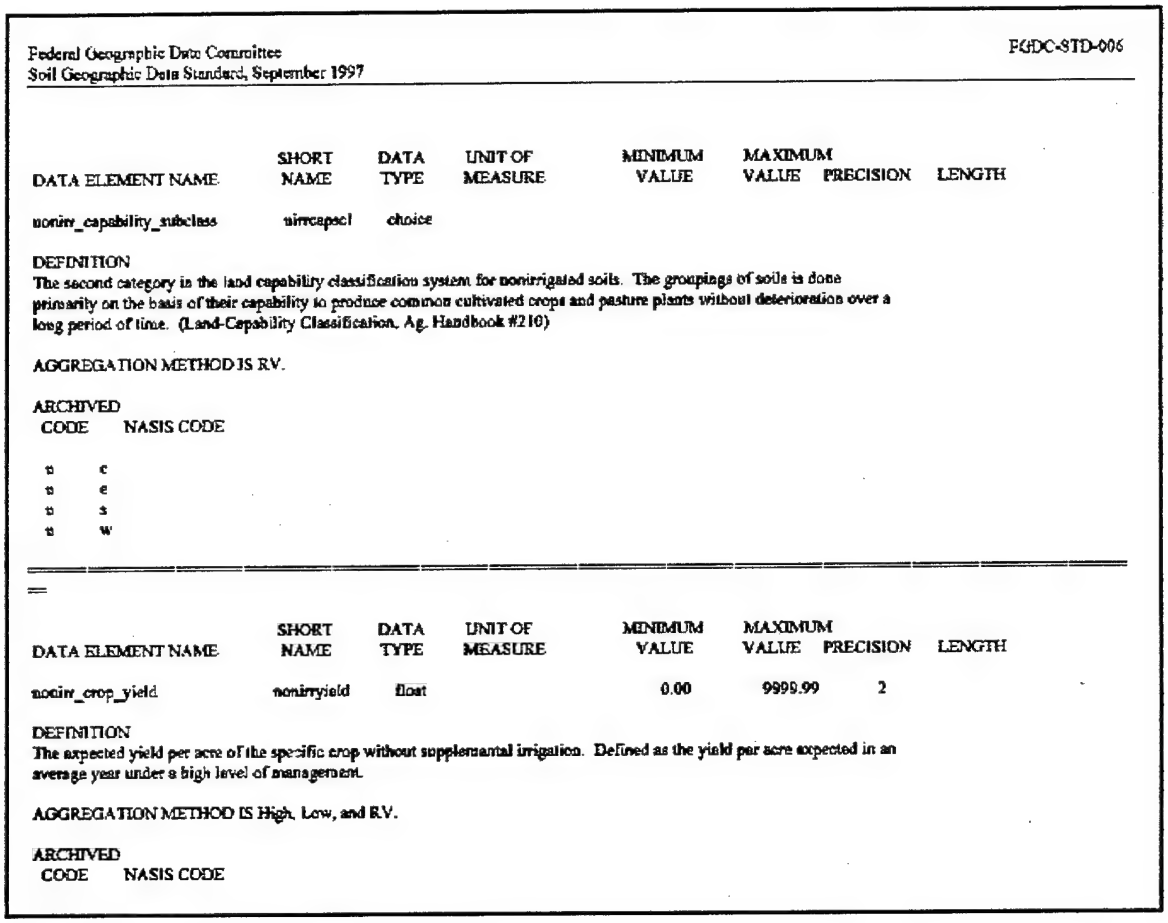
Standards Relevant to All Federal Agencies

Spatial Data Transfer Standard

The SDTS defines a protocol for transferring earth-referenced spatial data between dissimilar computer systems. The standard calls for a self-contained transfer in that it must include all components of spatial data, including the features, attributes, georeferencing, data quality information, data dictionary, and other supporting metadata. Development of the SDTS began in 1980 under the direction of the U.S. Geological Survey (USGS). In 1992, after 12 years of developing, reviewing, and testing, the resulting standard was approved as FIPS Publication 173, known as FIPSPUB 173-1, 1994. The FIPS version has been superceded by ANSI NCITS 320-1998, which was ratified by the American National Standards Institute (ANSI). Compliance with SDTS is now mandatory for Federal agencies.

Data Content Standards

The FGDC subcommittees and working groups are advancing numerous standards that define data models for various spatial data types. These models define the data types and standardize the names, definitions, ranges of values, and other characteristics of their attribute data. Examples of existing standards include addresses, environmental hazards, soil characteristics, shorelines, coastal and inland waterways, and vegetation. All data content standards follow overall standards for semantics. Figure 1 shows an example of data element definitions from the Soil Geographic Data Standard.



Federal Geographic Data Committee
Soil Geographic Data Standard, September 1997

FGDC-STD-006

DATA ELEMENT NAME	SHORT NAME	DATA TYPE	UNIT OF MEASURE	MINIMUM VALUE	MAXIMUM VALUE	PRECISION	LENGTH
nonirr_capability_subclass	nirrcapscl	choice					

DEFINITION
The second category in the land capability classification system for nonirrigated soils. The groupings of soils is done primarily on the basis of their capability to produce common cultivated crops and pasture plants without deterioration over a long period of time. (Land-Capability Classification, Ag. Handbook #210)

AGGREGATION METHOD IS RV.

ARCHIVED

CODE	NASIS CODE
n	c
n	e
n	s
n	w

DATA ELEMENT NAME	SHORT NAME	DATA TYPE	UNIT OF MEASURE	MINIMUM VALUE	MAXIMUM VALUE	PRECISION	LENGTH
nonirr_crop_yield	nonirryield	float		0.00	9999.99	2	

DEFINITION
The expected yield per acre of the specific crop without supplemental irrigation. Defined as the yield per acre expected in an average year under a high level of management.

AGGREGATION METHOD IS High, Low, and RV.

ARCHIVED

CODE	NASIS CODE

Figure 1. Examples of element definitions.

Content Standard for Digital Geospatial Metadata

The CSDGM is an FGDC standard-defining metadata (data to describe other data) for geospatial data. The CSDGM (FGDC 1998) states that the purpose of metadata is to:

1. Organize and maintain an organization's investment in data.
2. Enable prospective users to determine the data's fitness of use for an application.
3. Provide information to data catalogs and clearinghouses about data types and availability.
4. Provide information to aid data access and transfer.

The FGDC Content Standard details the type and organization of metadata needed to describe the content, quality, condition, and other characteristics of spatial data. The standard divides metadata into seven components:

1. Identification — name, developer, geographic extent, thematic types, currentness
2. Data quality — accuracy elements
3. Spatial data organization — spatial model, number of objects, encoding methods
4. Spatial reference coordinate systems, datums, conversion parameters

5. Entity and attribute information — definitions, content description, coding/ representation standards
6. Distribution — format, media, price, location for obtaining data
7. Metadata reference — developer, date compiled.

For each of these components, the FGDC standard formalizes an extensive set of elements and definitions to fully describe the seven components. Each element is specified as mandatory, mandatory if applicable, and optional.

The standard was approved in 1994. EO 12906 requires Federal agencies to use the standard to document data they produce as of 1995. The EO does not specify the means by which this information is organized in a computer system or in a data transfer, nor the means by which this information is transmitted, communicated, or presented to the user. Figure 2 shows example element definitions from the CSDGM.

National Standard for Spatial Data Accuracy

The NSSDA is an FGDC standard promoting a well-defined statistic (root mean square error [RMSE]) and testing methodology for positional accuracy. This standard is intended to apply to maps and geospatial data derived from sources such as aerial photographs, satellite imagery, or maps. Accuracy is reported in ground units. The testing methodology is a comparison of data set coordinate values with coordinate values from a higher accuracy source for points that represent features readily visible or recoverable from the ground. While this standard evaluates positional accuracy at points, it is intended to apply to geospatial data sets that contain point, vector, or raster spatial objects. Data content standards, such as FGDC Standards for Digital Orthoimagery and Digital Elevation Data, will incorporate the NSSDA for particular spatial object representations.

Standards Specific to the Department of Defense

Spatial Data Standard

The SDS is a comprehensive master and environmental planning data model for U.S. Air Force, Army, and Navy installations, as well as Corps of Engineers' Civil Works projects. The SDS not only defines terminology, data requirements and types, but also symbology to ensure standard map development. While the FGDC considers 11 classes of data, the SDS considers 24, including landform, geology, soil, cultural, transportation, utilities, and military operations. The

Computer-Aided Design and Drafting (CADD)/GIS Center is responsible for both developing the standard and advocating its use.

A design criteria for the SDS is that it follow and complement FGDC standards. Representatives from the CADD/GIS Center participate on FGDC subcommittees and working groups. Figure 3 shows an example specification.

Federal Geographic Data Committee FGDC-STD-001-1998
Content Standard for Digital Geospatial Metadata

Data Quality Information

2 Data Quality Information -- a general assessment of the quality of the data set. (Recommendations on information to be reported and tests to be performed are found in "Spatial Data Quality," which is chapter 3 of part 1 in Department of Commerce, 1992, Spatial Data Transfer Standard (SDTS) (Federal Information Processing Standard 173): Washington, Department of Commerce, National Institute of Standards and Technology.)
Type: compound
Short Name: dataqual

Data_Quality_Information =
0{Attribute_Accuracy}1 +
Logical_Consistency_Report +
Completeness_Report +
0{Positional_Accuracy}1 +
Lineage +
(Cloud_Cover)

Attribute_Accuracy =
Attribute_Accuracy_Report +
(1{Quantitative_Attribute_Accuracy_Assessment}n)

Quantitative_Attribute_Accuracy_Assessment =
Attribute_Accuracy_Value +
Attribute_Accuracy_Explanation

Positional_Accuracy =
0{Horizontal_Positional_Accuracy}1 +
0{Vertical_Positional_Accuracy}1

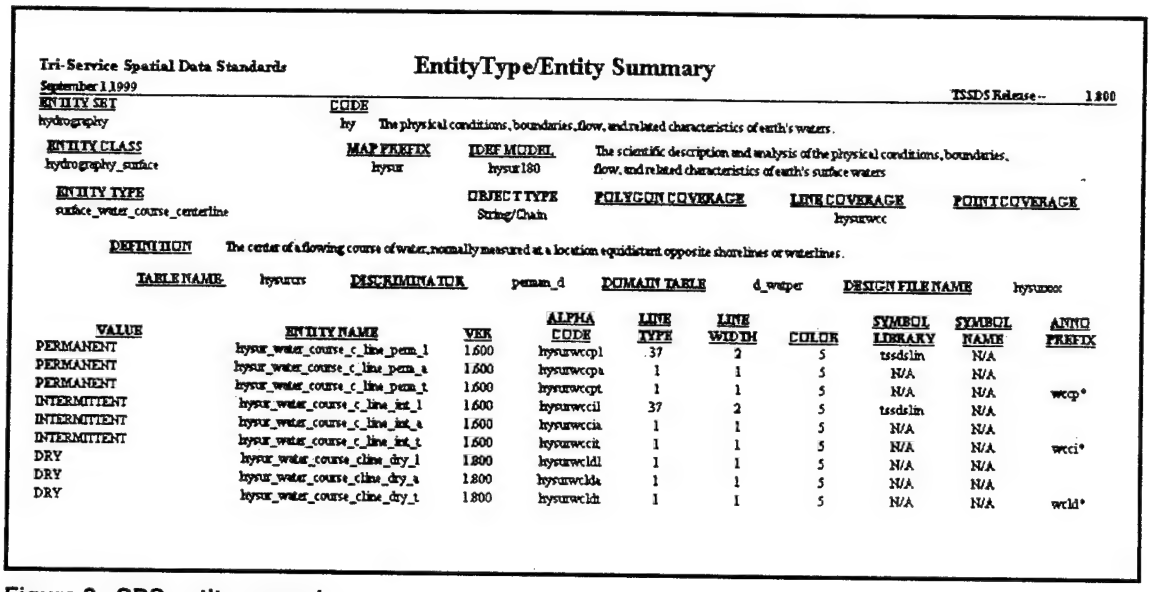
Horizontal_Positional_Accuracy =
Horizontal_Positional_Accuracy_Report +
(1{Quantitative_Horizontal_Positional_Accuracy_Assessment}n)

Quantitative_Horizontal_Positional_Accuracy_Assessment =
Horizontal_Positional_Accuracy_Value +
Horizontal_Positional_Accuracy_Explanation

Vertical_Positional_Accuracy =
Vertical_Positional_Accuracy_Report +
(1{Quantitative_Vertical_Positional_Accuracy_Assessment}n)

Quantitative_Vertical_Positional_Accuracy_Assessment =
Vertical_Positional_Accuracy_Value +
Vertical_Positional_Accuracy_Explanation

Figure 2. Element examples from CSDGM.



Tri-Service Spatial Data Standards

September 1 1999

EntityType/Entity Summary

TSSDS Release-- 1.800

ENTITY SET: hydrography
CODE: hy — The physical conditions, boundaries, flow, and related characteristics of earth's waters.

ENTITY CLASS: hydrography_surface
MAP PREFIX: hysur
IDEF MODEL: hysur180
The scientific description and analysis of the physical conditions, boundaries, flow, and related characteristics of earth's surface waters

ENTITY TYPE: surface_water_course_centerline
OBJECT TYPE: String/Chain
POLYGON COVERAGE:
LINE COVERAGE: hysurwcc
POINT COVERAGE:

DEFINITION: The center of a flowing course of water, normally measured at a location equidistant opposite shorelines or waterlines.

TABLE NAME: hysurcrs DISCRIMINATOR: perman_d DOMAIN TABLE: d_watper DESIGN FILE NAME: hysurccc

VALUE	ENTITY NAME	VER	ALPHA CODE	LINE TYPE	LINE WIDTH	COLOR	SYMBOL LIBRARY	SYMBOL NAME	ANNO PREFIX
PERMANENT	hysur_water_course_c_line_perm_l	1600	hysurwccpl	37	2	5	tssdslin	N/A	
PERMANENT	hysur_water_course_c_line_perm_a	1600	hysurwccpa	1	1	5	N/A	N/A	
PERMANENT	hysur_water_course_c_line_perm_t	1600	hysurwccpt	1	1	5	N/A	N/A	wccp*
INTERMITTENT	hysur_water_course_c_line_int_l	1600	hysurwccil	37	2	5	tssdslin	N/A	
INTERMITTENT	hysur_water_course_c_line_int_a	1600	hysurwccia	1	1	5	N/A	N/A	
INTERMITTENT	hysur_water_course_c_line_int_t	1600	hysurwccit	1	1	5	N/A	N/A	wcci*
DRY	hysur_water_course_cline_dry_l	1800	hysurwcldl	1	1	5	N/A	N/A	
DRY	hysur_water_course_cline_dry_a	1800	hysurwclda	1	1	5	N/A	N/A	
DRY	hysur_water_course_cline_dry_t	1800	hysurwcldt	1	1	5	N/A	N/A	wcld*

Figure 3. SDS entity example.

Elements of Geospatial Data Quality

Data quality is still an evolving concept in the geospatial research and application communities. Research consensus, however, is that data quality cannot be adequately described with a single component, and most concerned agencies have adopted standards incorporating the following five elements adopted in the SDTS (http://rmmcweb.cr.usgs.gov/public/nmpstds/sdts.html): lineage, positional accuracy, attribute accuracy, completeness, and logical consistency. The two final elements discussed here, semantic accuracy and temporal accuracy, are not components of the FGDC standard, but closely relate to or interact with other aspects of data quality.

Lineage

The lineage element of spatial data quality documents the history of the data set. It is not a quantitative indicator so much as a log of the data's life cycle. Lineage includes identification of the producer and descriptions of the events, parameters or assumptions, source observations or materials, compilation methods, conversions, transformations, or derivations in the process of developing the data set.

Lineage is useful to both the data producer and the data user. For the producer, it serves as a documentation mechanism and a production-tracking tool to help record and preserve the organization's investment in data development. For the user, lineage of acquired data sets can be useful in guiding additional data

development, assessing risk points in the development process, and providing base documentation for secondary or derived data sets.

Positional Accuracy

The positional accuracy element of data quality describes the degree of discrepancy between a data set's definition of the position of its objects and either the objects' actual positions (e.g., as measured in the field) or an accepted representation (e.g., another data set of recognized higher accuracy). Horizontal positional accuracy (x and y dimension) applies to all data sets, while vertical positional accuracy is limited to data sets that report a third (z) dimension. The positional accuracy element of the data quality report (1) describes the selected objects and method of evaluation and (2) provides a quantitative statistic representing likely nearness to true position (typically RMSE). These statistics are separately stated for horizontal and vertical accuracy.

Conceptually, determination of spatial accuracy is a simple comparison of the data set under review against another data set and a measurement of spatial discrepancies for selected entity objects. The ideal comparison is to the true state (i.e., the true location of the entity object on the earth's surface). An accepted comparison is to another data set of known higher quality. A third option is comparison to the source document from which the data set was derived, a process that generally only identifies error in the encoding process.

An error of positional accuracy can be introduced at numerous points in the development process. It can be random (e.g., a measurement error), systematic (e.g., a calibration error), or cumulative in nature. Recent research is looking at ways to measure spatial error (or accuracy) at each step in the development process. This measurement would allow developers to evaluate the relative costs of improving specific procedures versus their benefit in terms of accuracy improvements. Post-development application of this concept requires complete documentation of lineage. Regardless of whether this detail exists for a given data set, the approach could be adopted by the data user in determining cost versus benefit of quality improvement in secondary (derived) data sets.

Positional accuracy has been broken into absolute and relative components (U.S. Department of Interior 1990). Absolute positional accuracy is defined as how closely all the positions in a data layer match corresponding positions of the "true" features on the ground. Relative positional accuracy is a measure of how closely all positions in a data layer represent corresponding geometrical relationships on the ground. It reflects the consistency of accuracy for any position on a map with respect to any other position on that map.

The NSSDA has produced guidelines for determining the horizontal and vertical accuracy of the data. The draft conference report (1998) provides a worksheet for determining the RMSE and the NSSDA statistic, determined by multiplying the RMSE by the standard error at the 95 percent confidence level. RMSE is a function of the accuracy of the positional measuring or recording device and procedure. RMSE is described by empirical frequencies, means, and standard deviations of positional errors (Caspary and Scheuring 1993). While these standards quantify absolute accuracy, little insight is provided into the relative accuracy of the data (Stanislawski, Dewitt, and Shrestha 1996).

In vector-based GIS, the accuracy of a point is a fundamental issue. Historically, point accuracy has been used for line error analysis by the GIS community. Point error, however, can be used as a fundamental building block to construct models to analyze point, line, and polygon errors together. Leung and Yan (1998) proposed such an error model for point, line, and polygon features. It works by interpreting the minimum buffer width around the reference object that contains all the tested object, or vise versa. Their model simultaneously accounts for the *circular normal distribution model* for positional errors in points and the *epsilon band model* (with certain confidence limits) for errors in lines. This differs from other models that treat errors in point and line separately.

Goodchild and Hunter (1997) found that the epsilon band is very sensitive to outliers and sample size, which makes this method less robust, and hence, a less than ideal method as a QC measure. The assumptions used in the epsilon band model were that one knows what proportion of the mapped boundaries is within the limits of uncertainty imposed by the technology and what contain additional uncertainty. Skidmore and Turner (1992) assumed that the first form of uncertainty is known and focused on measuring the second. The method that Goodchild and Hunter (1997) proposed is non-parametric, looking at the problem of measuring positional uncertainty in the most general circumstance where nothing is known. They consider a buffer of width x around the reference source and compute the proportion of the tested source length that lies within the buffer. They also propose that this method could be applied to other forms of linear data (besides the coastline features they tested) and generalized to area and volume features.

Quantitative accuracy of line data can also be assessed by the buffer and overlay statistic (BOS) method. In this method, the approximate epsilon band is found first, then the average displacement is determined, and finally, a determination of the generalized level of completeness is made (Tveite and Langaas 1999). Two assumptions for this test are that all lines must exist in both data sets and the reference data must be significantly better. Once the lines are buffered and the

overlay made (with standard GIS operations such as BUFFER, UNION, and STATISTIC), one can calculate (1) the completeness of the data relative to another data set and (2) the miscoding of the attributes data.

Attribute Accuracy

Attributes are the non-spatial characteristics of an entity. Generally attributes are uniform across an entity, and serve to distinguish one object from another. Attribute values can be unstructured text descriptions (e.g., name = Lake Michigan, color = blue), nominal values (can be text or numeric, e.g., soil type, zip code, identification number), ordinal values (can be text or numeric, e.g., military rank, ordered classes), and interval (strictly numeric, embedded scale in the differences between values, e.g., elevation values).

The attribute accuracy element of data quality describes how well the assigned attribute values match the actual characteristics of the objects. The SDTS has slightly different requirements for interval attribute types as opposed to all other types, which are grouped together as nominal. Attribute accuracy for interval attributes must be a quantified assessment, a numerical estimate of expected discrepancies similar to positional accuracy. Attribute accuracy for nominal attributes can be assessed by either deduction, independent sampling, or independent polygon overlay. In addition to the assessment, this element of quality documents the test date, materials used, and descriptions of the test method for both types of attributes.

For remotely sensed data, one of the most common ways to represent the classification accuracy of data is with an error matrix. This is a square array of rows and columns expressing the number of sample units (i.e., points, polygons, or pixels) assigned to a particular category relative to the actual category as verified on the ground. Columns usually represent the reference data while rows represent the classification. The error matrix can then be used to describe and analyze the data with statistical techniques (Congalton 1991).

The simplest descriptive statistic derived from the error matrix is "overall accuracy." This is computed by dividing the total correct classifications (i.e., the sum of the major diagonal) by the total number of sample elements or observations in the error matrix. The accuracy of individual categories is computed in a similar way. Usually, the number of correct sample elements in a category is divided by the total number of sample elements of that category as derived from the reference data. This accuracy measure, the "producer's accuracy," indicates the probability of a reference sample element being correctly classified and is really a measure of omission error. If the total number of correct sample units in a

category is divided by the total number of elements classified in that category, this result is a measure of the reliability (the "user's accuracy"). This measure is indicative of the probability that the sample unit classified on the map correctly represents that category on the ground (Congalton 1991).

Other metrics derived from comparison matrices are *percent correct* (the sum of the diagonals divided by the total of all elements in the matrix). A change or comparison matrix (one representing the same area at different times) can be normalized to produce probabilities of change for dynamic modeling, or analyzed to determine if significant change has occurred. This method compares individual matrix cells between error matrices regardless of the number of samples. It is an iterative proportional fitting process that forces the rows and columns in a matrix to sum to one, which eliminates differences in sample size and allows for direct comparison of the individual matrix cell values regardless of the number of samples used to construct the matrix (Congalton 1991).

One important assumption is that the values used in the error matrix must be representative of the entire area mapped. Congalton (1988) examined sampling schemes used in generating the error matrices. Five sampling schemes were evaluated: simple random, stratified random, cluster, systematic, and stratified systematic unaligned. Each sampling scheme has advantages and disadvantages, greatly influenced by spatial autocorrelation (the pattern of the error) within the data set. Congalton (1988) found that, in his study area (agriculture), simple random sampling was the best.

While the error matrix is a good descriptive technique for spatial data, it is also the beginning for many analytical statistical techniques. Results from the error matrix can also be used as input into more advanced statistics. One statistic is discrete multivariate analyses, appropriate where data are binomially or multinomially distributed rather than normally distributed (in other words, when data are constrained into a few possible outcomes or types). A second statistic is *Kappa*, which calculates the percent correctness of a map and allows for comparison to other maps (Congalton 1991). Studies have found Kappa useful and credible for analyzing the relative strengths and weaknesses of two data sets (Greenland, Socher, and Thompson 1985).

Finn (1993) proposed the *average mutual information (AMI)* as an information theory measure of shared information. AMI measures a different aspect of the problem than either percent correct or Kappa, it measures correctness, quantifying the amount of information that one map contains about another map. Therefore, use of a combination of Kappa and AMI to assess error matrices is desirable

because of their different viewpoints in comparing classifications and potential for spotting mislabeling problems.

Completeness

Completeness is perhaps the most poorly defined element of data quality, with different standards using different terminology, often without clarification. The 1985 National Committee for Digital Cartographic Data Standards (NCDCDS) defines it as "an attribute describing the relationship between the objects represented in a database and the universe of all objects." The FGDC standard focuses on "information about omissions, selection criteria, generalization, definitions used, and other rules used to derive the data set."

A discussion by Brassel et al. (1995) provides a good breakdown of issues in the definition of completeness and role of the information with respect to fitness of use. Completeness, as a component of quality information, is defined as an indicator of whether each feature or entity is present in the data set, and whether all of its attributes are present. The completeness measured by the provider is a relative measure, comparing the data set's objects versus what it is intended to represent. Completeness for the user, in assessing the data set's fitness for use, must support consideration of not only the provider's completeness measure, but also whether the represented set of features or entities is compatible with the user's application requirements.

The standards for completeness assessment are all very general and do not contain formalized descriptions about how to measure the amount of missing information (Brassel et al. 1995). The CSDGM element for completeness is an unformatted text item, implying a statement about completeness rather than a formal assessment. The SDTS provides some guidance in terms of information that should be provided in a completeness report:

- Selection criteria, definitions, mapping rules
- Deviations from standard definitions
- Discrepancy between the objects in the data set and the set of real world objects.

The discrepancy can be a description or a quantitative measure. A description based on expert knowledge of the data producer is considered acceptable. A quantitative measure is preferred because it provides an assessment that can be concretely compared with a similar assessment for another data set. But the success of a quantitative assessment of completeness hinges on whether the selection criteria, definitions and rules, and deviations from standard definitions

can be determined. In other words, without knowing what the data set is supposed to describe, it is impossible to assess completeness.

Two measures of completeness are needed because of two possible types of errors: omission and commission. Errors of omission occur when a feature in the control data does not have a corresponding feature in the test data. Errors of commission occur when a feature in the test data does not have a corresponding feature in the control data (Table 1).

Table 1. Feature completeness.

	Control Data	
Test Data	**Present**	**Absent**
Present	Correct	Error of Commission
Absent	Error of Omission	Correct

Logical Consistency

This element describes the structural integrity of a data set. This integrity concept, relevant to any type of data, is concerned with assurance that identified constraints on data keys, attribute domains, and key and attribute interrelationships are observed. In addition to these issues regarding data values, logical consistency for spatial data is also concerned with geometric or topological consistency (Kainz 1995).

Documented procedures for consistency assessment are all very general and the standards do not contain formalized descriptions about how to measure the fidelity of the relationships. Like completeness, the CSDGM element for consistency is an unformatted text item. The SDTS provides some guidance in terms of information requirements for a consistency report:

- Selection of valid values and constraints
- Graphic rules for spatial reference method – e.g., prohibitions on intersections, nodes, minimum/maximum length or area
- For graphic rules, indicate 100 percent correction or detail remaining errors by case.

Unlike the other elements of spatial data quality, consistency assessment does not require a control data set from a source of higher quality. The consistency of the data set is determined through comparison to a descriptive model, approved procedures, or real world constraints.

The FGDC standard is open-ended with regard to specification of logical consistency. This study considers consistency similar to completeness, and identifies the tests performed, rules adopted, and a quantification of the test results.

Semantic Accuracy

Semantics refers to the meanings of words or symbols. Semantic errors in language arise, for example, when someone uses an incorrect definition for a word. Semantic errors most frequently arise among different subgroups of a population (e.g., in the interpretation of teen-age slang by parents). Another, more relevant example, would be possible differences in interpretation of the word "grassland" between an ecologist and a rancher or by two different groups using different words when referring to the same actual object (e.g., "bush" vs. *Cornus racemosa*).

For spatial data, semantic accuracy is defined as "the quality with which geographical objects are described in accordance with the selected model" (Salgé 1995). It is assessed by measuring the number of errors of commission or omission in naming attributes in the data set relative to a model and specification. In other words, semantic accuracy is measured by the relative number of labels that occur in a data set that should not occur according to the data developer's specification or the number that do not occur, but should.

Semantic accuracy is an element not currently required by the FGDC but can be closely related to the concept of completeness (although that refers as much to the proper inclusion of entity objects as to their attribute meaning).

Temporal Accuracy

Temporal accuracy refers to the relationship between the temporal characteristics of the database representation versus the "real-world" entity, or its "datedness." Temporal information is more than just part of the description of the lineage of a data set. It is an attribute at multiple levels in the spatial data definition. It can apply to an overall object, describing when the object became relevant to the selected model (when a road was built). It can also apply to an attribute of an object, describing when a change in classification occurred (the road was widened to four lanes). It can also apply as an attribute of the collection of the data (roads were last surveyed in 1984).

Because of its diversity of relevance, temporal information interacts with all other aspects of data quality. As a separate element of data quality, it is a

relatively young concept. It is not a component of the FGDC standard, and there are no methodologies for quantified measures to represent it (Guptill 1995).

Implications

The literature provides a history of consideration of the aspects of spatial data errors and is somewhat uneven in its coverage of the various components. To make this information useful in documenting and maintaining ITAM databases, there is a need to determine and describe concrete methods of assessment, documentation, and research into their efficacy.

Specific methods for assessing the various dimensions of spatial data quality are not well documented or reported in the literature. While a substantial effort has been directed toward data documentation through metadata standards, it is certainly arguable that relatively little information exists to allow GIS managers to assess and report data quality information, particularly for existing or legacy data sets. Most commercial software packages provide little to no functionality for quality assessment although some of the general analysis tools can be used for part of an assessment. Therefore, a principle purpose of this study is to explore and test methods for assessment, to include a preliminary report based on the case study of the Fort Hood ITAM data base.

3 Assessment Methodology

The proposed methodology is based on the elements of quality assessment described in Chapter 2. From a decision-making perspective, this methodology initiates consideration of options for data management once an assessment has been completed. That is, while it is inherently useful to know the existing quality of a data set, it is also useful to understand the nature of the requirements for improving the quality of a data set with respect to the marginal utility of improved information. Therefore, the proposed methodology consists of the following stages:

1. Determine the data set's lineage.
2. Determine the scope of real world phenomena intended to be represented by the data set through metadata on parent material or through best available expert knowledge.
3. Identify a control or reference data set.
4. Determine an appropriate sample set of objects for testing.
5. Assess the data set's accuracy.
   - Positional
   - Attribute
6. Assess consistency.
7. Assess completeness.
8. Provide documentation.
9. Evaluate requirements/utility of improving current data set.

Factors in proposing a particular method or approach include feasibility, cost (e.g., labor), and marginal improvement in quality. It should not be presumed that these factors have been thoroughly examined when reviewing an existing data set. The purpose of using Fort Hood as a case study is in part to assess the operational requirements for conducting an assessment. Determining the reliability of these factors requires testing against multiple data sets.

Terms Used in Proposed Methodology

The vocabulary associated with geographic information is diverse, often reflecting particular implementations of systems, or the nomenclature used by software

vendors to describe their implementation of geographic data. For consistency's sake, this report uses a set of terms defined here.

Object: A digital representation of a particular instance of an entity (e.g., a road vector, or stream, or a point rural drop location).

Data Set: A digital collection of objects (e.g., the set of roads, the set of streams). A data set typically represents entities having the same structure and description.

Data Base: A digital collection of data sets, with associated methods for querying data sets.

Test Data: A set of objects drawn from a data set to be used to estimate the quality of the data set.

Entity: A real world phenomena (e.g., road, stream, rural drop location).

Control Data: A set of objects drawn from a data set or collected in the field that serves as the standard of comparison for the test data.

Determine the Data Set's Lineage

A detailed inspection of the actual spatial data is conducted to help determine lineage, identify gaps in the existing data, clarify relationships between data sets, and to assist in the selection of data sets for QA/QC testing and the development of control data.

Steps used in data review consist of:

1. Description of the basic data set (entity type, number of objects, spatial domain, spatial data organization, spatial reference).
2. Derivation of lineage information (source(s), processes used in development of the data set).
3. Identification of relationships between data sets for the same entity types (if applicable).
4. Development of a formal statement of the universe intended to be represented in the data set (e.g., surface drainage channels with drainage areas in excess of 10 acres, stream crossings approved and maintained by Range Control).

Resources used to inspect existing data sets include existing metadata or institution-developed documentation, software log/documentation files, existing knowledge of database managers, and data file inspection (inferring metadata from observable characteristics, or letting the data "speak for itself"). These resources are clearly in the order of preference as they represent the most direct information regarding the data set (although not necessarily the most accurate).

In many cases existing metadata or institutional knowledge are not available to document the characteristics of a data set. Determination of data set characteristics therefore may rely heavily on inferences based on comparison between the data set and possible parent materials. Factors that may be used to infer lineage include spatial extent and distribution of geographic objects, attributes and domains used in the data sets, or topological structure of the data.

Spatial extent of a particular data set may be an indication of the data's source, especially if it coincides with boundaries typical of sources such as USGS quadrangles that are partitioned based on defined, regular geographic boundaries (Figure 4). Matches in the spatial distribution of the objects may indicate the data's completeness, if entities of the same type are known to exist elsewhere in the study area.

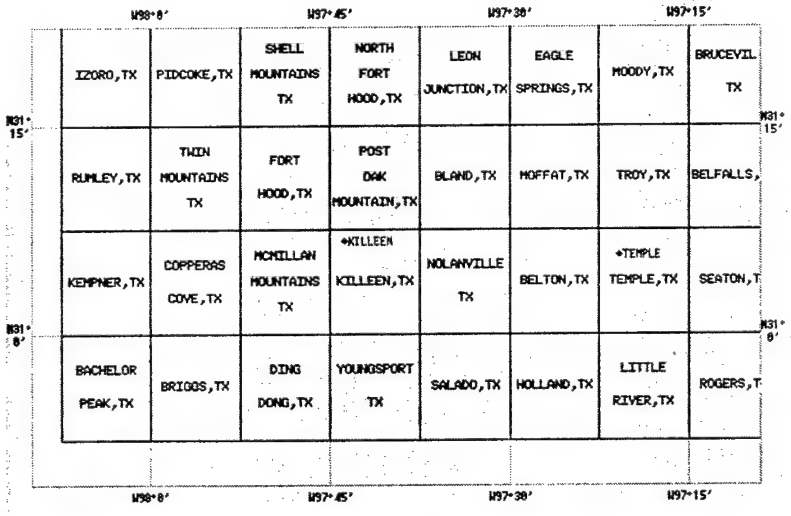


Figure 4. Quadrangle partitioning of data.

Attributes and their domains can indicate data source, such as USGS Digital Line Graphics (DLG) or U.S. Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER) data (Figure 5). Attributes may also be used as an indication of the data's completeness and of the types of assessments that can or should be performed.

Given the identification of candidate source material, lineage can be confirmed through comparison of geometry/topology (amount of detail, connectivity, redundancy, or error) and spatial and aspatial evaluations (redundancy, logical consistency). Appendix A contains more information on creating secondary control data.

Chapter 6: Data Dictionary

Record Type 1---Complete Chain Basin Data Record

Field	BV	Fmt	Type	Beg	End	Len	Description
RT	No	L	A	1	1	1	Record Type
VERSION	No	L	N	2	5	4	Version Number
TLID	No	R	N	6	15	10	Tiger/Line ID, Permanent Record Number
.							
TRUSTR	Yes	L	A	128	128	1	American Indian Trust Land Flag, Current Right
CENSUS1	Yes	L	A	129	129	1	Census Use 1
CENSUS2	Yes	L	A	130	130	1	Census Use 2

BV (Blank Value):
Yes = Blank value may occur here; No = Blank value should not occur here
Fmt:
L = Left-justified (numeric fields have leading zeros and may be interpreted as character data)
R = Right-justified (numeric fields do not have leading zeros and may be interpreted as integer data)
Type:
A—Alphanumeric, N = Numeric

Figure 5. Excerpt from U.S. Census Bureau TIGER Data Dictionary.

Identify Control Data

Quality assessment of geospatial data is an exercise in relative performance. Because the data set is by definition a representation of real world phenomena, thus representing a simplification, evaluation can only be accomplished by comparing the result against the intended model. Performance of an assessment requires comparison of the test data against some reference.

A source of higher quality data typically serves as a comparative model for any accuracy test. A data set is considered of higher quality than the test data if it has one or more of the following characteristics:

1. Represents more detail (is at a larger scale).
2. Used more rigorous data quality assurance procedures for data collection.
3. Used higher quality instrumentation.
4. Comprises a more recent measurement.
5. Consists of direct observations/measurements in the field.

The use of higher quality data as a reference set introduces something of a conundrum for the user. If a higher quality data set is available, why not discard the original and replace it with the higher quality data? Typically the reason is based on the completeness of the reference data set. It may contain a subset of the objects at higher resolution or accuracy. It may not represent all attributes. The cost of complete development may not be feasible. Finally, the quality may exceed that required for the application of the data, rendering the additional accuracy useless.

The NSSDA provides a detailed statistical and test methodology for estimating positional accuracy. Typically, only a sample of objects from the reference data set is necessary to assess the accuracy of an entire data set. The standard states that "a minimum of 20 check points" should be tested (FGDC 1998). At a 95 percent confidence level, one point out of the 20 can fall outside the estimated error. The standard recommends that the sample be distributed so that 20 percent of the points fall within each quadrant of the area to be tested. This rule may be ignored when the control data exist for only a portion of the test area or if objects are distributed unevenly.

By setting such a small sample size, the standard is encouraging data developers to measure and document the accuracy of their data sets. But a key point in the documentation should not be overlooked; the sample needs to "reflect the geographic area of interest and the distribution of error in the data set" (FGDC 1998). This qualification suggests that the reliability of the test and the resulting error estimation depends on how well the sample objects tested represent the entire data set. The 20-point sample size becomes inappropriate if there is reason to suspect that the data set may not be homogeneous in terms of accuracy. This is highly likely if the data were developed over time or with unspecified development control procedures, or if the data set contains multiple entity types or describes significantly different areas.

Standard statistical procedures direct that the sample size be determined as a function of the target confidence level of the result, the size of the population being tested, and the expected variation in the test factor. For a homogeneous data set, where development procedures are known, 20 points may be a reasonable sample. However, if there is reason to doubt the consistency of the data set, then a larger sample, such as 40 or 60 points, should be selected.

Assess the Data Set's Quality

The four elements that should be independently evaluated to completely describe a data set's quality are: positional and attribute accuracy, completeness, and logical consistency. Each of these is in some way influenced by the data set's lineage. An understanding of the original source and process used to produce the data set should influence the selection of appropriate tests. As indicated in the earlier discussion of the elements of QA/QC, inaccuracies in one element may have an impact on the accuracy of another element. The specific tests presented for each element are intended to isolate the measurement of inaccuracy of that element.

QA procedures that have been studied most are positional accuracy of GIS data as measured by the spatial accuracy of the data (geographic placement) and the topological relationships (network connectivity). While some studies still rely on visual and programmatic review procedures (Gaertner 1993), various means of calculating error are available.

The assessment methods are discussed in the following sections. A suggested procedure for conducting each assessment is presented, along with a suggested metadata record entry for that type of accuracy assessment.

Positional Accuracy

The positional accuracy element of data quality summarizes the difference between the objects' true locations and their locations as defined in the digital data set. The "true location" is a somewhat elusive point, given that any measurement of location will introduce some error. Thus the true location is taken to be ideally the most accurate possible measure of the location, which conventionally is a recently collected real world position reading taken from a high-order survey or with a high-precision global positioning satellite (GPS) receiver. If this true location is unobtainable, a proxy can be used as the control measure. In the case of a proxy, positional accuracy needs to describe both the error of the control data versus the test data and the error measure of the control itself.

A quantitative measure for the positional accuracy of point data has been adopted as a standard by the NSSDA. However, measuring positional error for entities that are represented as lines, areas, or volumes is more problematic. The positional error for these entities cannot be identified at one specific location, but can occur at an infinite number of locations within the entity's path or extent. There is much research discussing alternative approaches to identifying, measuring, and reporting these compound positional errors. No one method has

been recognized as a standard, and many have yet to be defined in a way that can be easily implemented with a GIS or other tools. The various approaches were evaluated on how easily the resulting statistic could be interpreted, and how readily they could be performed with currently available GIS software. One approach has been selected for each spatial reference method (point, line, polygon).

The CSDGM requires both a quantitative measure of positional accuracy and a free text description. The sections that follow include a suggested approach to documenting the quantitative measure, per the NSSDA. Unfortunately this recommendation cannot be used in the metadata if the content standard is explicitly followed (separate items for the numeric statistic and the statistic's description). The statistic itself should not be considered sufficient to describe accuracy, and the free text description for the metadata should include a detailed explanation of how the test was performed.

Point feature method.

The standardized accuracy assessment method uses the RMSE positional accuracy measure. This statistic is advanced in the NSSDA (Spatial Data Accuracy Handbook 1998). The test is based on the identification of well-defined points in the data set. It is the only test appropriate for entity types that are represented as points (e.g., discrete entities such as traffic lights, telephone booths, or utility access locations).

RMSE is the square root of the average of the set of squared differences between the coordinate values for the test data and the control data. The statistic represents the largest expected error from ground position. The calculation is very well-defined and easy to implement in a spreadsheet or other calculation engine.

Positional accuracy is measured independently in the horizontal and vertical directions. The horizontal RMSE incorporates error in both the x and y directions, representing a radial measure of positional error in ground units. The vertical RMSE is a unidirectional measure that assumes x and y are equal for the comparison points.

The task that is most critical in ensuring the reliability of this test is the correct matching of the control and test points. With well-described data that can be identified by a key attribute, matching can be handled with an aspatial join. When correct key attributes do not exist in the test or control data set, objects could be tentatively matched using a spatial join. For well-distributed points this spatial matching is probably accurate. However, if points are densely

spaced and positional accuracy is low, the matches would have to be reviewed and confirmed based on the tester's knowledge of the control and test data sets.

The procedure for assessing horizontal positional accuracy consists of the following steps:

1. Collect x and y position measurements for the point objects in the control and test data sets.
2. Match the control points to the appropriate points in the test data set.
3. For the matched points, calculate the radial RMSE:
   $(\Sigma ((\text{control x} - \text{test x})^2 + (\text{control y} - \text{test y})^2)) / \text{number of matched points})^{1/2}$
4. Adjust the RMSE for a 95 percent confidence interval:
   RMSE * 1.7308

To assess vertical positional accuracy:

1. Collect the z value for the point objects in the control and test data sets. These do not necessarily have to be the same objects used for horizontal positional accuracy assessment.
2. Match the control points to the appropriate points in the test data set.
3. For the matched points, calculate the vertical RMSE:
   $((\Sigma (\text{control z} - \text{test z})^2) / \text{number of matched points})^{1/2}$
4. Adjust the RMSE for a 95 percent confidence interval:
   RMSE * 1.9600

A typical statement to include in the metadata for reporting the results of positional accuracy assessment using RMSE is:

Tested ___ (unit of measure) horizontal accuracy at 95% confidence level

Tested ___ (unit of measure) vertical accuracy at 95% confidence level

The unit of measure is usually the same as the map units for the data set. In addition to the statistic, the metadata should describe the methodology used to derive the statistic, including the control source and the number or percentage of points tested.

Linear features.

The NSSDA advocates using RMSE for linear entity types by basically deconstructing the objects into a series of point comparisons. This approach is acceptable when a set of specific points can be reliably identified in both the test and control data. Suitable well-defined points include intersections of objects within the same data set, or between the test objects and objects for some other entity

types. A typical example is a roads data set, where coordinate locations for right-angle road centerline intersections are used for positional error estimation of the road network. Restricting assessment to such points may not capture errors in the geometry of such data (e.g., the alignment of roads between intersections).

The point series approach is not always possible with complex objects, for several reasons. Streams entity types provide a good example of data that may be difficult to measure with confidence using the point series approach. Stream confluences would be a logical point location to use for comparisons. As illustrated in the Fort Hood study described in Chapter 4, however, they may be hard to access or identify from control sources. Streams could also be evaluated based on their intersection with another entity type such as roads, but this introduces the possibility that the error measure considers not just error in the stream data but also error in the roads data. Finally, selected stream intersection points may have greater positional accuracy than is true for the data set in total. Because these points are generally more identifiable, less error may have been introduced in the data production process by identifying them versus the linear component in between (Goodchild and Hunter 1997).

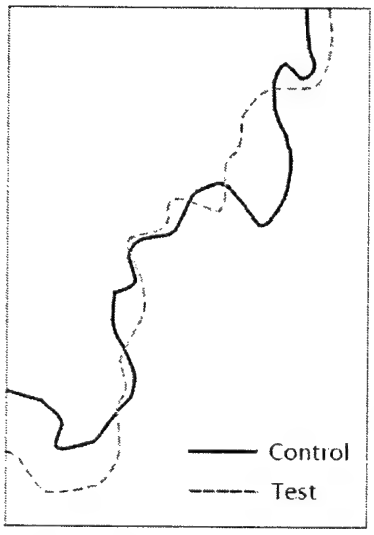
An alternative proposed by Goodchild and Hunter (1997) uses a distance buffering method to derive an error estimate that considers the entire object. Figure 6 illustrates the process. A full buffer (buffer on both sides of the object) is placed around the control data. The test data are evaluated to determine what proportion is contained within the buffer distance (in the example, a 40-meter buffer around the control data includes 31.57 percent of the test data). As the distance from the control increases, so does the percentage of included test data.

When the calculation is performed for a series of distances, the approach can provide a percentile distribution of accuracy (Figure 6, part C). The 95 percent percentile is the distance at which 95 percent of the test data falls within the calculated buffer of the control data. The error estimator is the distance buffer, and the percentile is a kind of confidence level.

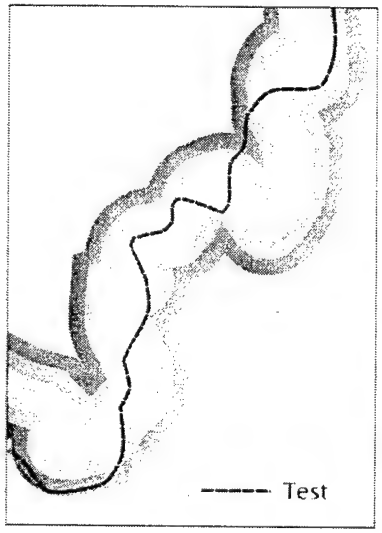
To assess horizontal positional accuracy:

1. Reduce the control and test data sets to only those objects that exist in both sets.
2. Determine the target percentile or a range of percentiles for accuracy documentation. Selection of a range of percentiles provides greater information to potential users than a single statistic.
3. Estimate the test distance(s) based on available indicators of positional accuracy such as source scale, development methods, or known accuracy of similar data sets.
4. Construct the buffer(s) around the control objects.

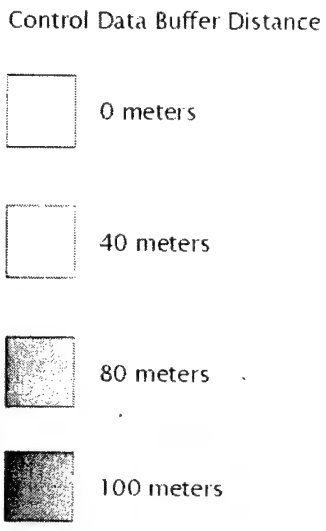
5. For the matched lines, extract the test lines that fall within the buffer(s) and compute percentage:
   (test length of line within buffer) / Σ (test length of all lines)
6. If the calculated percentage does not match the desired percentiles, repeat the test from step 3, with adjusted buffer widths.



A. Overlay of Control and Test Data of a Stream Segment



B. Process of Clipping the Test Data From a Buffer Distance



C. Results From the Buffer/Clip Processes

Buffer Size	Frequency	Sum of Lengths	% of Line Within Buffer
0	345	273426.59	0
40	1206	86322.67	31.57
80	742	172341.88	63.03
100	402	263005.34	96.19

Figure 6. Distance buffers around entities.

This approach does not include a method for measuring vertical positional accuracy.

A logical statement to include in the metadata for reporting the results of positional accuracy assessment using line buffering is:

Tested ___ (unit of measure) horizontal accuracy at (percentile) percentile

The unit of measure is usually the same as the map units for the data set. The test statement can be repeated for different percentiles. In addition to the statistic, the metadata should describe the methodology used to derive the statistic, including the control source and the number or percentage of line objects tested.

Polygon features.

The NSSDA does not discuss positional accuracy testing methods for polygon entities. The assumption is that the same rules for using RMSE for linear entity types can be extended to polygons, which are basically enclosed linear entities. The issues for applying a point series approach are the same as those for linear entities. Assuming that a set of specific points can be reliably identified in both the test and control data, however, an RMSE statistic may be sufficient. Suitable points might include a common intersection point of polygon boundaries, or a well-defined feature on which the boundary is based, such as a road edge.

When specific points cannot be identified, the buffering test described above for linear entities can be used as an appropriate test. To measure both the overstatement and understatement of the polygon, buffers should be placed around the linear object defining its area. The test is performed in the same manner, yielding a statistic representing positional error in the same unit of measure as the data set.

An alternative approach to representing polygon error is the correlation statistic Kappa (Greenland, Socher, and Thompson 1985). The statistic represents the proportion of agreement of the test versus control data set over and above chance agreement. The method is intuitive in that it matches how a reviewer might visually compare two data sets, using a boolean overlay approach to compare the areas and measure the geometric accuracy of the lines defining the polygons.

Figure 7 illustrates the methodology. In Part A, the solid line (C) represents the polygon object for the control data and the dashed line (T) represents the polygon object for the test data. The result of an overlay procedure is displayed in Figure 7, Part B. Four distinct classifications of areas are derived from the overlay operation:

- Areas located within both *Control* and *Test*
- Areas located within *Control*, but outside *Test*
- Areas located within *Test*, but outside *Control*
- Areas located outside both *Control* and *Test*.

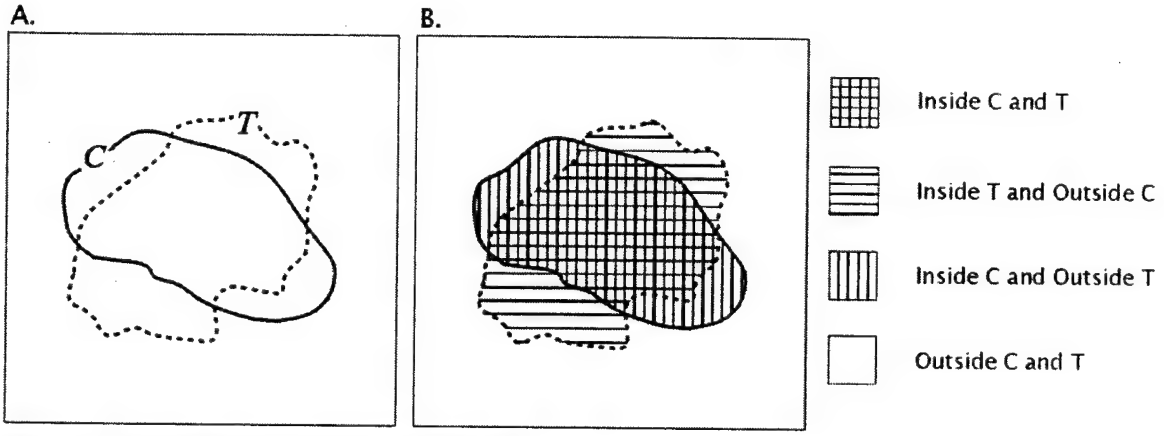


Figure 7. Methodology of the Kappa Statistic.

Table 2 summarizes the agreements and disagreements between the data sets according to the four classifications above. The values (P_{ii}, P_{io}, P_{oi}, P_{oo}) are fractions of the total area in the data set. Total area can be any area measure that exceeds the area of all the polygons being tested.

Table 2. Summary of agreements/disagreements for the Kappa.

		Classified by *Test* data		
		IN	OUT	
Classified by *Control* data	IN	P_{ii}	P_{io}	P_{i+}
	OUT	P_{oi}	P_{oo}	P_{o+}
		P_{+i}	P_{+o}	$P_{++} = 1$

(Reproduced with permission, the American Society for Photogrammetry and Remote Sensing. Greenland, Socher, and Thompson 1985. Original table has been modified for this report.)

To assess horizontal positional accuracy:

1. Reduce the control and test data sets to only those objects that exist in both sets.
2. To simplify the output and interpretation of the overlay, add an attribute item to each of the data sets (e.g., code_t in the test data, code_c in the control data). For all entity polygons, assign the attribute one value (e.g., "in"). For the polygon within each data set that defines total area, assign a different value ("out").
3. Perform an overlay operation to create a new data set that divides the polygons into the four classifications (inside C and T, etc.).
4. Summarize the areas by classification type:
   P_{ii} = Σ area where (code_t="in" & code_c="in") / total area
   P_{io} = Σ area where (code_t="in" & code_c="out") / total area
   P_{oi} = Σ area where (code_t="out" & code_c="in") / total area
   P_{oo} = Σ area where (code_t="out" & code_c="out") / total area
   Depending on the tool used to perform the overlay, the actual calculations for determining total area and P_{oo} may need to be adjusted. (Chapter 4 provides an example for determining Kappa using ARC/INFO.)

5. Calculate the percentage of agreement between the test and control data:
   $P_a = P_{ii+} P_{oo}$
6. Calculate the expected fraction of agreement:
   $P_e = (P_{+i} * P_{i+}) + (P_{+o} * P_{o+})$
7. Calculate the Kappa statistic:
   $K = (P_a - P_e) / (1 - P_e)$

This approach does not include a method for measuring vertical positional accuracy.

A logical statement to include in the metadata for reporting the results of positional accuracy assessment using the Kappa statistic is:

> Tested Kappa = ___ (proportion of agreement over and above chance agreement)

The percentage of agreement (P_a) can be reported in the metadata, in addition to or as an alternative to the Kappa statistic. Regardless of the quantitative measures used, the metadata should describe the methodology for how the statistics were derived, including the control source and the number or percentage of polygon objects tested.

Attribute Accuracy

The attribute accuracy element of data quality summarizes the differences between the objects' categorization as defined in the test data set and their true categorization. The "true categorization" is observed ideally from a recently conducted field survey. There are many situations, however, in which a true categorization cannot be observed, ranging from physical inaccessibility to the non-observable nature of some attributes. For these instances, a proxy — a categorization that has been recorded or produced independently of the test data set — must be obtained. When a proxy is used, attribute accuracy needs to describe both the error measure of the control data compared to the test data and the error measure of the control itself.

Currently no standards have been formally adopted for performing a specific attribute accuracy test. The content standard for metadata, however, requires both a quantitative measure of attribute accuracy and a test methodology description. The implication is that there are formulas, statistics, and methods that would provide a measure appropriate for comparison.

The literature on attribute accuracy is not nearly as extensive as that for positional accuracy, but there are several methods that seem to be accepted. These methods share some common characteristics with positional accuracy testing, which makes sense when both location and description are seen as different kinds of attributes of an entity. Regardless, some of the same tests for positional accuracy can be applied in attribute accuracy assessment.

The decision about what test to apply is based on how critical the accuracy of the attribute is and what type it is. The simplest test, percent correctly classified, identifies the fraction of the tested data set known to contain erroneous categorizations, and serves loosely as a probability of miscategorization for any entity. This measure is probably sufficient for simple descriptive label attributes such as names. Other attributes that identify physical characteristics may require a greater understanding of their accuracy because they may influence subsequent analyses. These attributes can be separated into either nominal or interval/ratio types. Nominal attributes are simple classes, such as road type. Interval/ratio attributes are those values representing a uniform measurement scale, such as temperature or slope.

The task that is most critical in ensuring the reliability of attribute accuracy testing is the identification of an appropriate sample. The selected entities should be well-described data that can be identified by a key attribute, to eliminate matching based on location and isolate the issues of attribute error from issues of positional error. For nominal attribute types, it may be appropriate to use a stratified sampling approach. This approach would consider the number of categories and the difficulty of correctly assigning an object to a particular category in the selection of objects to test. A basic assumption before applying any attribute test is that standards for semantics have been defined and applied, for example "road" is always represented as "rd" (errors of this type are identified with consistency testing).

Descriptive label.

Descriptive label attributes are generally free text data for which it is impossible to calculate differences, means, or standard deviations. Generally an error exists when the test value does not match the control value for a particular object, and all errors contribute equally to the accuracy of the attribute classification. The test for descriptive labels is the percent correctly classified, a simple average.

To assess attribute accuracy for descriptive label attributes:

1. Select appropriate objects for comparison using a simple random sample.
2. Join the two sets of objects based on a key identifier.

3. For the matched objects, calculate the percent correctly classified:
   (Count (control attr == test attr)) / number of matched objects)

A logical statement to include in the metadata for reporting the results of attribute accuracy assessment using an averaging statistic is:

Tested PCC = ___ (percent correctly classified)

Nominal attributes.

Nominal attributes may be text strings or numeric data used to assign entities to discrete classifications. The number of classes is finite, and there is only one representation for each class ("first" is a class but not also "1st"). Like descriptive labels, a nominal attribute error exists when the test value does not match the control value for a particular entity, and all errors contribute equally to the accuracy of the attribute classification.

A PCC measure can be used for nominal data. An alternative measure that can be used is the Kappa statistic, which is essentially the same statistic described for positional accuracy. Here, however, it is extended to consider more than two categories and to be performed aspatially. Instead of a matrix with two rows and columns, the matrix contains as many rows and columns as entity classifications (Table 3).

Table 3. Determining attribute accuracy with Kappa.

		Classified by *Test* data				
		A	**B**	**..**	**N**	
Classified by *Control* data	**A**	P_{aa}	P_{ab}	$P_{a..}$	P_{an}	P_{a+}
	B	P_{ba}	P_{bb}	$P_{b..}$	P_{bn}	P_{b+}
	..	$P_{..a}$	$P_{..b}$	$P_{....}$	$P_{..n}$	$P_{..+}$
	N	P_{na}	P_{nb}	$P_{n..}$	P_{nn}	P_{n+}
		P_{+a}	P_{+b}	$P_{+..}$	P_{+n}	Total

To assess attribute accuracy for nominal attributes:

1. Select appropriate objects for comparison using a simple random sample.
2. Join the two sets of objects based on a key identifier.
3. To simplify the output and interpretation of the join, add a matrix classification item to the join result. Set the matrix item to a value indicating the type of attribute match per the matrix table above.
4. Count the number of objects in each attribute match type and calculate the fraction of the total to fill in the row/column values for the matrix.

5. Calculate the percentage of agreement between the test and control data:
   $P_a = \Sigma P_{nn}$
6. Calculate the "expected" fraction of agreement:
   $P_e = \Sigma (P_{+n} * P_{n+})$
7. Calculate the Kappa statistic:
   $K = (P_a - P_e) / (1 - P_e)$

A logical statement to include in the metadata for reporting the results of attribute accuracy assessment using the Kappa statistic is:

Tested Kappa = ___ (proportion of agreement over and above chance agreement)

The Kappa statistic's percent of agreement (PA) component is identical to the percent correctly classified (PCC) statistic. This can be reported in the metadata, in addition to or as an alternative to the Kappa statistic.

Interval/ratio attributes.

Interval and ratio attributes are strictly numeric data. Values within the scale are related such that the error associated with a misclassification from 1 to 2 is less than the error for a misclassification from 1 to 10. Attribute accuracy assessment for these types of data is simple because the data can be presumed to fall within the framework of a normal error distribution (Goodchild 1995). The RMSE is an appropriate calculation here; it incorporates issues of mean error and standard deviation in one measure that is in the same units as the data being analyzed.

To assess attribute accuracy for interval/ratio data:
1. Select appropriate objects for comparison using a simple random sample.
2. Join the two sets of objects based on a key identifier.
3. For the matched objects, calculate the RMSE:
   $(\Sigma (\text{control attr} - \text{test attr})^2) / \text{number of matched objects})^{1/2}$

A typical statement to include in the metadata for reporting the results of attribute accuracy assessment using RMSE is:

Tested ___ (unit of measure) RMSE

The unit of measure is the same as the units for the attribute.

Per attribute testing and metadata.

Testing should be performed for each attribute in the data set. The sample set and control source can be the same or different for each attribute. The metadata allows for multiple instances of a quantitative statistic, so each quantitative description should include the name of the attribute. In addition to the statistics, the metadata should describe the methodology used to derive the statistics, including the control source and the number or percentage of objects tested. There is only one descriptive report field, so the metadata methods entry should clearly distinguish each attribute tested.

Completeness

Like accuracy, completeness can be divided into two components: entity completeness and attribute completeness. Entity completeness refers to the exhaustiveness of the data set in terms of the entity type it is intended to represent. Attribute completeness refers to the exhaustiveness of the attributes — the physical and descriptive information provided about each entity. Both components imply a comparison against the selection criteria for, or standard intended to be met by, the data set.

Entity completeness.

Entity completeness indicates the degree to which all entities that exist in the field (e.g., lakes or buildings) have a matching object in the data set. The task that is most critical in ensuring the reliability of entity completeness testing is the identification of an appropriate sample short of a complete field survey. A logical approach is a cluster sampling technique to systematically survey a set of areas and determine the number of entities that should be represented. This approach would consider the extents of the test data set and any variability in the difficulty of correctly identifying an entity. Field collection as a source of higher accuracy for completeness testing may be cost-prohibitive, as real world systematic sampling of an area is time-consuming. A proxy — identification of entities that have been recorded or produced independently of the test data set — must be obtained. Review of orthoimagery is a good source for control data when the entities are readily identifiable. When a proxy is used, completeness needs to describe both the error measure of the control data compared to the test data and the error measure of the control (e.g., the accuracy/resolution of the orthoimage) itself.

A simple quantitative measure appropriate for entity completeness is percent of agreement. This ratio is the number of objects in the test data consistent with

the control data compared to the number of objects in the control data. The selected objects should be well-described data that can be identified by a key attribute, to eliminate matching based on location and isolate the issues of completeness from issues of positional error.

To assess entity completeness:

1. Select appropriate areas to inspect for comparison.
2. Identify the objects in that area from the control data set.
3. Identify the objects in that area from the test data set.
4. Join the two sets of objects based on a key identifier.
5. Calculate the percent of agreement (PA):
   (Number of matched entities / Number of entities in the control data)

A logical statement to include in the metadata for reporting the results of completeness assessment using the PA statistic is:

Tested PA (entities) = ___ (percent of agreement)

The ratio is always a value between 0 and 1. This value is ensured by restricting the numerator to matched objects rather than all objects. It is possible for the test data to overstate the set of objects by including objects that are not confirmed in the control data. This is identified by another calculation, the percent of excess, which is the number of objects in the test data inconsistent with the control data compared to the number of objects in the test data.

To assess entity overstatement:

1. Using the results of the join from the omissions test, select the objects from the test data set that did not match entities from the control data set.
2. Calculate the percent of excess (PE):
   (Number of unmatched test objects / Number of objects in the test data)

A logical statement to include in the metadata for reporting the results of completeness assessment using the PE statistic is:

Tested PE (entities) = ___ (percent of excess)

Attribute completeness.

The task that is most critical in ensuring the reliability of attribute completeness testing is the identification of the appropriate attributes. Ideally these attributes are identified in a standard model for the entity type being represented. The FGDC is advancing standards for many entity types such as cadastral, cul-

tural, and demographic, hydrographic, biological, and engineering data. SDS is a recommended standard for military installation spatial data. Numerous standards exist, and no single standard is required for assessing completeness.

No guidance for attribute completeness is provided by the FGDC. In fact, the breakdown between entity and attribute completeness is not required. But the same assessments can apply to both. The quantitative PA measure can be modified, so the ratio is determined as the number of attributes in the test data set in agreement with the standard compared to the number of attributes in the standard. The quantitative measure is meaningless, however, without a description of what standard was applied.

To assess attribute completeness:

1. Identify the standard attributes — the set of physical and descriptive fields necessary to describe the entity.
2. Compare the standard to the attributes defined in the test data, identifying matches.
3. Calculate the percent of agreement (PA):
   (Number of matched attributes / Number of standard attributes)

A logical statement to include in the metadata for reporting the results of completeness assessment using the PA statistic is:

Tested PA (attributes) = ___ (percent of agreement)

Consistency

Consistency as a general term deals with logical rules of the structure and relationships between data in a database. Spatial data is a specialized database describing entity objects with two important components: a descriptive component of attributes, and a physical component of the graphic elements and their relationships. Both components should be independently reviewed for consistency.

Attribute consistency.

Each attribute should be independently tested according to its expected constraints; this is what the FGDC identifies as a "test of valid values." For any attribute test, a logical measure is a simple ratio indicating the percent that complied with the test versus the number tested.

To assess attribute consistency:

1. Identify the permissible values for an attribute.

2. Select a random sample of objects to test.
3. Evaluate the object's value for the attribute being tested against the set of permissible values (domain). Any value not in the domain generates an error.
4. Calculate the percent of agreement (PA):
   (Sample size – number of errors / Sample size)

A logical statement to include in the metadata for reporting the results of consistency assessment using the PA statistic is:

Tested PA (attribute a) = ___ (percent of agreement)

Physical consistency.

Ideally issues of physical consistency are identified and corrected during the data development process. But in the event of incomplete metadata, methods for assessing and correcting physical consistency may be critical to prepare a data set for a particular application. Each data set generally contains only one entity type and one spatial reference method (point, line, polygon). Both of these characteristics will determine the appropriate constraints and test approaches.

Physical constraints applicable to point data are issues of location: do neighboring points violate minimum distance requirements? Tests to verify this constraint can be automated with the GIS software, and violations of this type of error may be related to issues of positional accuracy.

Physical constraints applicable to line and polygon data consider what provides a complete and accurate indication of each object and how the objects relate to each other. Figure 8 illustrates many of these:

1. Are all objects completely described graphically?
2. Do any objects contain overshoots or undershoots?
3. Do objects intersect only where intended?
4. Do any objects exist twice?
5. Are any objects too close?
6. Are any polygons too small (sliver)?
7. Do any polygons overlap?

Different tests may be applied to address these questions. The quality report should contain a description of the tests applied or a reference to explanatory documentation. The SDTS requires that all inconsistencies be corrected or that the quality report identify any remaining physical consistency errors case by case.

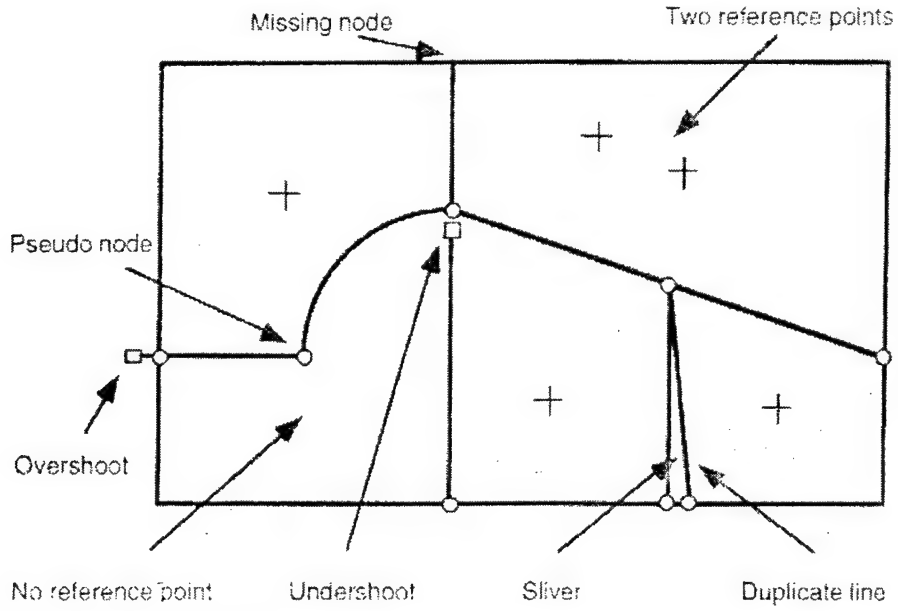


Figure 8. Geometric errors in the Topology of GIS data.

(Reprinted from Kainz 1995, with permission from Elsevier Science.)

GIS software tools often contain automated procedures to assess and correct some simple topological relationships, such as:

1. Nodes not separated by a minimum distance are merged.
2. Lines intersect at nodes according to an exact case or tolerance.
3. Dangling lines meet a minimum length requirement.
4. Cycles of lines and nodes are consistent around polygons. Or, alternatively, cycles of lines and polygons are consistent around nodes.
5. Inner rings embed consistently in enclosing polygons.

If automated software tools are used, the quality report should identify the software (name and version) and any parameter settings that would alter its processing.

Develop Metadata

Metadata is most accurate and complete if it is developed concurrently with the spatial data and updated with each modification to the spatial data. As the example study illustrates, it is very difficult to discern critical lineage information in a post-development investigation.

The exhaustive detail of the metadata standard itself can be intimidating. Numerous tools now available, however, can help an organization manage the development of the metadata. The Topographic Engineering Center (Alexandria,

VA) sponsored the development of one such software tool, CORPSMET95. This tool guides the metadata developer through the requirements of the standard. It uses icons to indicate which elements are mandatory and when sufficient elements have been entered to complete a logical section.

The FGDC defines the format for the final output of the metadata as a formatted ASCII* report using the terminology and organization defined by the standard. CORPSMET95 supports two outputs: a proprietary format (a "*.gen" file) and the ASCII format (a "*.met" file). The *.gen file is the working file. It can be produced or updated at any time in the metadata development process and stores all entered information, complete or incomplete. CORPSMET95 will only produce a *.met file with completed mandatory information.

Some guidelines for using CORPSMET95 for metadata development are:

1. Create one or several "template" metadata files containing standard information (organization references, distribution information, and spatial reference and organization). Use this as a starting point for new metadata.
2. Maintain a metadata file for each spatial data set; use the data set's name for the metadata file name (for example, the data set "roads" will have a metadata file "roads.gen").
3. Store the metadata in the same location as the spatial data set.
4. Treat the metadata as a required component of that data set; any time the data is distributed to another person or location, deliver the ASCII-formatted metadata file with the data set.
5. Incorporate development and update of metadata into the organization's methodologies for spatial data management.
6. Since CORPSMET95 is available to any user at no cost, include the delivery of metadata as a requirement in all contracts involving the development or analysis of spatial data.

CORPSMET95 version 1.2 was used to develop metadata throughout this project. While a systematic evaluation of CORPSMET95 was not performed, the following two issues were identified:

1. Since the tool is not integrated with the GIS tools used for development, analysis, display, and reporting of the spatial data, initiation and management of the metadata is at the discretion of the user.

* ASCII – American Standard Code for Information Interchange.

2. The content standard for metadata has and will continue to change, with more explicit identification of elements, terminology, and standards. The tool does not allow for the end user to incorporate these changes. A decision needs to be made to continue to update and monitor use of the tool or to evaluate alternatives now available through software vendors.

4 Example Application: Selected Fort Hood Data

Examination of Fort Hood Data

To illustrate and evaluate the assessment methodology, researchers used the Fort Hood (Figure 9) ITAM database. Fort Hood's data is typical of much spatial data: data sets within the database were developed independently based on a particular need at a particular time, and metadata are scarce. Current users of this data have varied knowledge of the development methods and any related quality assurance or control mechanisms that exist or may have been applied.

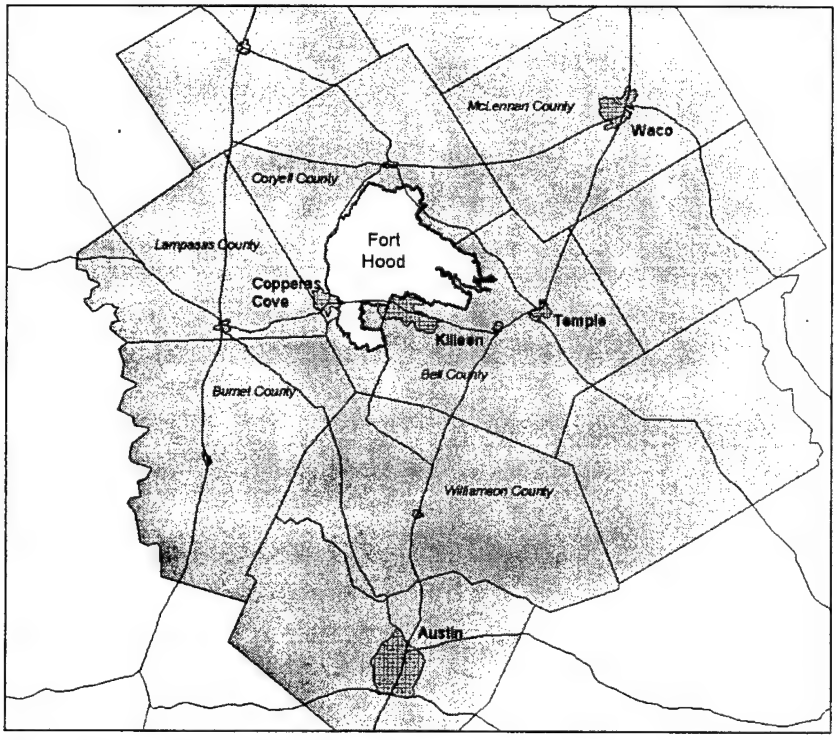


Figure 9. Location of Fort Hood, Texas.

Discussions between Fort Hood, the Construction Engineering Research Laboratory (CERL), and the Geographic Modeling Systems (GMS) Laboratory (University of Illinois at Urbana/Champaign) personnel resulted in a prioritized list of data sets to be addressed by the project:

1. installation boundary
2. training area boundary
3. roads
4. surface hydrology
5. stream crossings
6. pipeline crossings
7. rural drop locations.

Description of Source Data

The majority of the Fort Hood spatial data used for this study was received in two transfers in February 1999. The first transfer was a compact disc referencing over 600 MB of various geographic data. A second transfer of 37 data sets was delivered electronically by File Transfer Protocol (FTP). Several limited transfers occurred between March and June 1999, delivering data sets that were being updated as the QA project got underway. The data inspection focused on the files received by FTP as these data were specifically selected from the substantial library of data sets at Fort Hood. Files from the initial transfer were added to the inspection if they seemed particularly relevant or contained data not represented in the primary data sets.

All data were delivered in the Environmental Systems Research Institute's (ESRI's) ArcView* product's proprietary format, called "shape" files. Table 4 identifies, by entity type, the number of data sets received and their method of spatial reference.

Fort Hood ITAM personnel were in the process of developing internal documentation for their data sets; this information was delivered with the data. This documentation included, for each data set, a one-line description of the data and its origins, a notation of source (ITAM, DPW, Archeology, GRASS), and date of last update. The information was insufficient, however, to complete the lineage component for the metadata. Also, there were often several data sets for one en-

* In this report, specific references to products are a result of the contract specifications and should not be taken as an endorsement by CERL or the contractors.

tity type and little documentation describing their differences. A detailed inspection of the data was undertaken in an attempt to discover any metadata residing in the data sets themselves.

Table 4. Fort Hood file transfers.

Entity Type	Number of Data Sets	Spatial Reference Method
Installation boundary	3 (Feb 1999); 1 (May 1999)	Polyline; polygon
Training area boundaries	5 (Feb 1999); 2 (Jun 1999)	Polyline & polygon; polygon
Roads	9 (Feb 1999)	Polyline
Surface hydrology	10 (Feb 1999); 1 (Jun 1999)	Polyline
Stream crossings	3 (Feb 1999)	Point
Pipeline crossings	2 (Feb 1999); 1 (May 1999)	Line
Rural drop locations	1 (May 1999)	Point

Inspection of Original Datasets

The inspection organized data by entity type. All of the inspection work was done using ArcView facilities, from simple data display and visual review to comparative operations such as queries and joins. A detailed example of the process and documentation from the inspection is included in Appendix B.

Each data set was first reviewed independently to determine the spatial reference information, the number and distribution of objects, the descriptive attributes and attribute domains. The data shared a common spatial reference system, and generally the same extent. Some data sets extended beyond the limits of the installation, often coinciding with USGS 7.5-min quadrangles. In other instances data had been limited to the extent of an obsolete description of the installation boundary, or to specific subareas within the installation.

The most commonly occurring attributes were fields typically used in other software systems (e.g., ARC/INFO topologic attributes or computer-aided design [CAD] layer attributes). This indicated that the data sets originated in other systems and were converted to ArcView format. Features containing zero values for these items were assumed to indicate that the data set had been updated in ArcView since the conversion. Descriptive attributes (e.g., names, type characteristics, or physical properties) were rarely present. Some entity type attributes were implied by the organization of data into separate themes or files, such as data sets for small versus large streams.

In general the content and organization of the data sets did not provide conclusive information about their source, intermediate processing tasks performed on the data, age of the data, or even a clear definition of the set of entities being represented. USGS DLG data was identified as the likely parent of a road and a

stream data set because of their extents and attributes. The current release of this data was used to produce detailed comparisons for confirming the parentage and identifying any modifications.

The second phase of the review consisted of comparisons between data sets. These were intended to discern parentage, a definition of the set of entities being represented, and any redundancies or gaps between data sets within an entity type. Automated techniques included attribute and spatial join, but the results of these often had to be visually reviewed to ensure correctness.

Most GIS software offers many tools that can be used for aspatial and spatial evaluations. The ArcView query tool for aspatial evaluation is very robust, supporting filtering, comparisons to specific values, comparisons to values in or derived from other attributes within the same data set, and complex combinations of these comparisons. Attribute analysis across data sets is limited to exact matching on a single field, although this limit can be circumvented by creating new fields representing the combined values of two or more attributes.

ArcView also contains several tools for spatial evaluations. Entities from one data set can be identified based on its topological relationship (contained within, intersect with, within distance of) to entities in another set. A spatial join operation supports the matching of entities across data sets based on their spatial proximity:

1. Point features match based on "nearest," and a distance measure between matched source and destination objects is calculated. The reviewer can filter out exact spatial matches (distance of zero), and compare attribute values of other matched objects to identify objects that have been relocated versus objects that are unique instances.
2. Polyline features match based on precise location and geometry. The described rule for a match is when a source object "is part of" an object from the destination data set. Objects that share the same location and geometry for all or part of their extents will match as long as all points from the source objects exist in the destination objects.
3. Polygon features match with less precision than polyline and point features. The match rule is when a source object "is inside" a destination object. Thus, the entire area encompassed by the source polygon must be within the area encompassed by the destination polygon. Depending on the join order (which table is the source and which the destination) the match results can be different. The method for finding duplicate polygons is to identify those polygons that match regardless of join order.

The point spatial join should have been an especially useful operation for the numerous stream-crossing data sets. Unfortunately, most of these had no consistent identification code or attributes, so it was impossible to determine whether objects that matched by location were distinct objects or the same object with different geography.

The spatial join operation for polyline data sets proved very useful for examining the many road and stream data sets. Spatial joins across the various stream data sets (separate data sets representing all streams, large streams only, small streams only, with differing extents) showed which sets had been made from a common parent. Varying the source and destination data sets for join operations helped identify which objects had been modified, and pointed out some areas where data could be better organized and maintained.

The spatial join between the data sets roads and roadmajall found no matches, but a detailed visual inspection showed that the objects in roadmajall were a subset of those in roads. Thus, both data sets originated from the USGS 7.5 DLGs. However, they must have migrated to ArcView using different transformation processes, resulting in slight differences in their coordinate positions.

The spatial join for polygon data sets was most useful for inspection of Fort Hood's water bodies (dams) data sets. The other area data sets, training areas and installation boundary, were represented by both polygon and polyline spatial types and could only be compared in a detailed visual inspection.

Results of Inspection

Table 5 provides a simplified summary of the inspection results. Unfortunately, there were many data sets for which parentage and currency could not be determined. Both the written documentation and the communicated institutional knowledge were incomplete in the concrete identification of the original sources, their scale and accuracy, the development and maintenance processes undertaken, and the currency of the data represented. Parentage was determined for only two of the data sets. In many instances data sets were related to each other, but an ultimate source was never confirmed.

The inspection process was generally unsuccessful in determining lineage, a key component for QC and assessment. However, it was extremely useful for gaining familiarity with the data set to be assessed. It also highlighted a significant need for GIS tools to incorporate metadata as an integral component in the management of spatial data.

Table 5. Data set inspection summary.

Entity Type / Data Set	Spatial Type	# of Entities	Presumed parent	Presumed Currentness	Unique	Recent Update	Extents	Descriptive Attributes	Selected for Testing
Installation boundary									
1954prop	Polyline	2203	Unknown	Unknown	No	No	NA	None	No
Boundary	Polyline	31	Unknown, smaller extents than universe	Unknown	No	No	NA	None	No
Universe	Polyline	50	Modified boundary	Unknown	No	Yes	NA	None	No
Universe	Polygon	1	Recently updated universe polyline using 1997 aerial imagery as reference, converted to polygon.	May 1999	No	Yes	NA	None	Yes
Training areas									
Livefire	Polyline	39	Selected objects from an old version of trnafull	Unknown	No	No	NA	One (implied)	No
Pd94	Polyline	17	Selected objects from an old version of trnareas	Unknown	No	No	NA	One (implied)	No
Trnareas	Polyline	290	Unknown, CAD converted	Unknown	No	No	Equal	None	No
Trnafull	Polyline	389	Unknown, ARC/INFO converted, recently updated using 1997 aerial imagery as reference	May 1999	No	Yes	Equal	None	No
Trnafullp	Polygon	140	Unknown, GRASS converted	Unknown	No	No	Equal	One	No
Trnareas_poly	Polygon	75	Modified trnafull, selected areas, converted to polygons	June 1999	No	Yes	Equal	One	No
Trnaful_poly	Polygon	140	Modified trnafull, converted to polygons	June 1999	No	Yes	Equal	Two	Yes
Roads									
Highways	Polyline	67	Unknown, GRASS converted	Unknown	No	No	Greater than	One	No
Roadmainold	Polyline	752	Objects of roadmajall,	Unknown	No	No	Equal	None	No

Entity Type / Data Set	Spatial Type	# of Entities	Presumed parent	Presumed Currentness	Unique	Recent Update	Extents	Descriptive Attributes	Selected for Testing
			clipped to boundary						
Roadmajall	Polyline	2729	Selected objects from USGS 7.5 DLG, ARC/INFO converted (for roadmajall parent entry)	1985	No	No	Greater than	None	Yes
Roads	Polyline	16944	USGS 7.5 DLG	1985	No	No	Greater than	Four	Yes
Roads_detailed	Polyline	76,326	Unknown, CAD converted	Unknown	No	No	Equal	One (implied)	No
Roads_imp	Polyline	600	Selected objects from roadmajall, some modifications	1997	No	No	Equal	One (implied)	No
Roads_lram	Polyline	166	Manually digitized from unreferenced source	Unknown	No	No	Less than	None	No
Roads_unpaved	Polyline	21869	Unknown, may be CAD converted	Unknown	No	No	Equal	One (implied)	No
Roadssec	Polyline	140	Selected objects from roadmajall, some modifications	1997	No	No	Equal	One (implied)	No
Surface hydrology/Water bodies									
Riverall	Polyline	2698	Unknown (likely shares common parent with rivers)	Unknown	No	No	Greater than	None	Yes
Riverlg	Polyline	640	Selected objects of riverall, clipped to boundary, some modifications	Unknown	No	No	Less than	One (implied)	No
Riverlgall	Polyline	284	Selected objects of riverall, (greater extents, fewer entities than riverlg)	Unknown	No	No	Greater than	One (implied)	No
Rivers (DLG)	Polyline	540	USGS 7.5 DLG	1985	No	No	Greater than	One	No
Rivers	Polyline	1712	CAD (likely shares common parent with riverall)	Unknown	No	No	Equal	One	No
Riversm	Polyline	1268	Selected objects of riverall, clipped to boundary, some	Unknown	No	No	Less than	One (implied)	No

Entity Type / Data Set	Spatial Type	# of Entities	Presumed parent	Presumed Currentness	Unique	Recent Update	Extents	Descriptive Attributes	Selected for Testing
			modifications						
Alldam97	Polygon	377	Multiple sources	Unknown	No	No	Greater than	Two (1% complete)	No
Alldam99bnd	Polygon	218	Modified alldam97, clipped to universe	May 1999 (partial)	No	Yes	Equal	Two,	Yes
Lakes (DLG)	Polyline	187	Selected objects of rivers (DLG)	1985	No	No	Greater than	None	No
Lakes	Polygon	23	Unknown	Unknown	No	No	Greater than	None	No
Ponds	Point	155	Unknown	Unknown	No	No	Equal	None	No
Stream Crossings									
Lowwaterxing	Point	13	Unknown (ITAM received Oct. 1998)	Unknown	No	No	Equal	One (implied)	Yes
Strmxing	Point	20	Manually digitized from unreviewed source	Unknown	No	No	Equal	One (implied)	Yes
Strmxpro	Point	74	Unknown	Unknown	No	No	NA	None	No
Pipeline Crossings									
Pipeline	Polyline	68	Modified DLG	Unknown	Yes	No	Greater than	Unknown	No
Pipelbnd	Polyline	27	Objects of pipeline clipped to boundary, lengths recalculated	Unknown	NA	No	Less than	Unknown	No
Pipexing	Polyline	21	Manually digitized from unreviewed source	Unknown	Yes	No	Equal	Unknown	Yes
Rural drop locations									
Teledrop	Point	59	Manually digitized from unreviewed source	Unknown	Yes	No	Equal	Unknown	Yes

Planning the Assessment

The information uncovered by the inspection provided the criteria by which to select the data sets for continued testing, editing/replacement, and metadata development. The data sets representing rural drop locations and pipeline crossings were selected without question because they were unique to their entity type. All other data sets were evaluated and several selected based on the following factors:

- Specifically identified or recently updated by ITAM personnel
- Confirmed against a reputable source
- Spatially as extensive or more extensive than the installation
- Included descriptive attributes.

The Table 5 inspection summary includes columns for these factors and an indicator of which data sets were selected for testing.

The key component for ensuring the reliability of all the accuracy tests was the selection of a source of higher accuracy. This selection was largely driven by the source and content of the data sets to be evaluated. The lack of information about the origins of the Fort Hood data, and the lack of publicly available data for many of the entity types, made the choice simple. The control data had to be the highest quality data available to ensure that it was more accurate than the unidentified test data. For the Fort Hood data, this meant field collection or interpretation of recent imagery. For several data sets, field collection was the only possible source of control data. Imagery could not be used as control data for the installation boundary and training areas data since the most current imagery was used for the recent update. The rural drop locations would be indistinguishable in any imagery (Figure 10). Table 6 summarizes the assessments that were planned for the Fort Hood data sets.

The majority of the assessment effort focused on positional accuracy. Attribute accuracy checking was eliminated first because of a lack of attributes, and second because the attributes that were represented could not be conclusively defined (for example, there was no identification of the rule distinguishing between "small" and "large" rivers). Consistency checking, applicable to both geometry and attributes, was limited to topological consistency checking for linear entities such as stream and road networks.

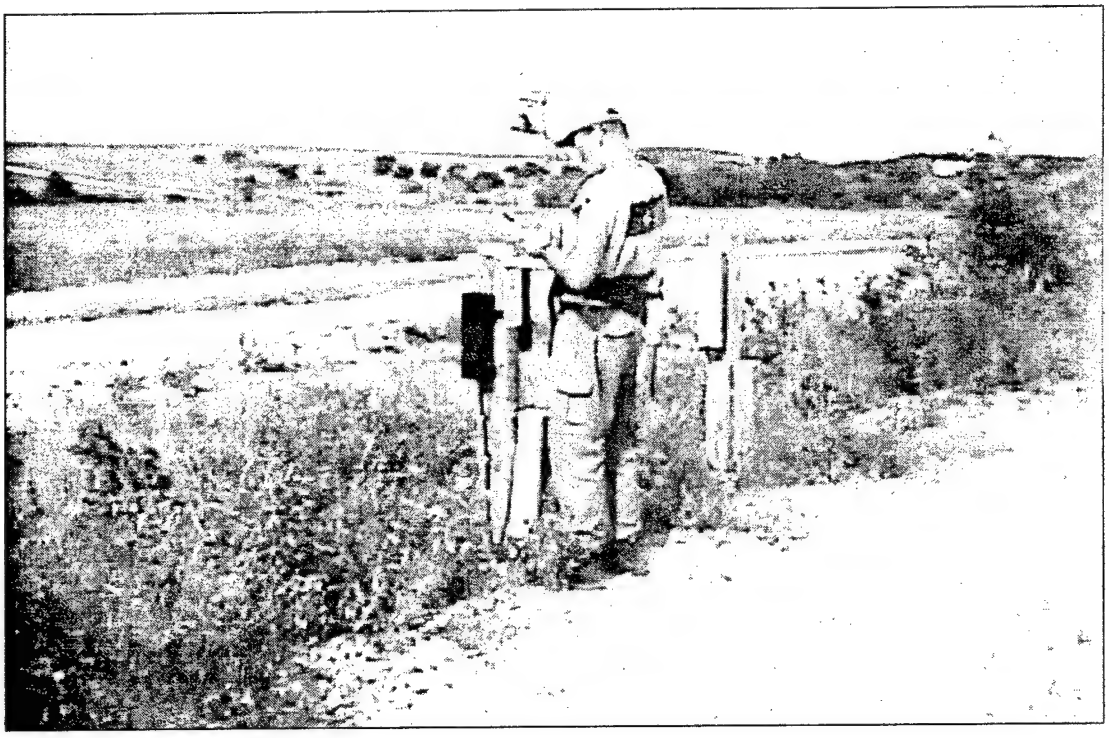


Figure 10. Teledrop site.

Table 6. Planned data set assessments.

Data Set	Control Data Sources	Positional Accuracy	Attribute Accuracy	Consistency	Completeness
Universe	Field only	Yes	No	No	No
Trnaful_poly	Field only	Yes	No		No
Roadmajall	Imagery	Yes	Yes	Geometry	Yes
Roads	Field, Imagery	Yes	No	Geometry	No
Riverall	Field, Imagery	Yes	No	Geometry	
Alldam99bnd	Field, Imagery	Yes	No	Geometry	Yes
Strmxing / Lowwaterxing	Field, Imagery	Yes	No	No	No
Pipexing	Field, Imagery	Yes	No	No	Yes
Teledrop	Field only	Yes	No	No	No

The ability to perform completeness assessments was limited by a number of issues. For area entities such as the installation and training area boundaries, the lack of a precise definition of their boundaries made completeness assessment impossible. For discrete entities such as lines and points, completeness required systematic surveying of a sample area. A field survey is costly, but imagery resolution limited the data sets to entity types that could be clearly recognized. Identification of stream crossings and tank trails from the imagery suffered from definitional issues — at what point did a worn patch of land become a sanctioned tank trail (Figure 11)? Only roads (*roadmajall*), water bodies (*alldamm99bnd*), and pipeline crossings (*pipexing*) could be reliably tested for completeness.

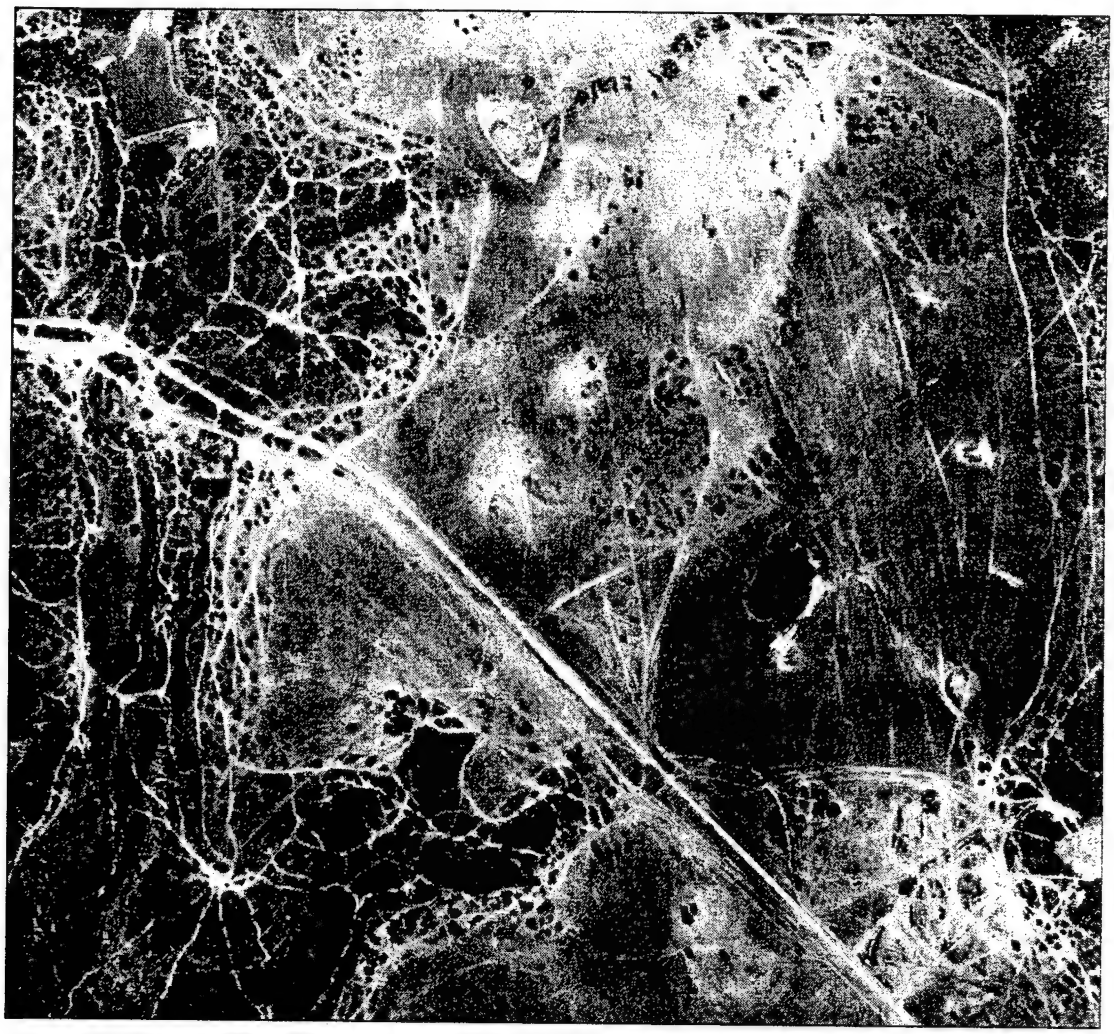


Figure 11. Tank trails.

Control Data Development

Control data sets were developed to assess the positional accuracy and completeness of the seven Fort Hood data sets examined in this report. To efficiently produce an accurate control data set, a number of strategies were examined. The appropriate methodology varied depending on the feature type (i.e., point, line, or polygon), time constraints, extent, and ease of collection for a particular geographic entity. Two methods of data collection were exercised: (1) field collection and (2) digitizing from digital media. The scope of the project and the extensive size of the installation meant that only a sampled selection of objects from each data set was necessary. The goal was to generate enough samples for each data set to have an accurate representation of the entire population. Generalizations could then be made detailing the overall quality of the test data. Different sampling techniques were implemented for both methods of data collection.

Field Collection

Preparation.

The use of GPS receivers is becoming a widely adopted approach to check for positional accuracy of an entity object. A sampling strategy was used to estimate the overall accuracy of entities due to time and resource constraints. While sampling provides a valid method for assessing data quality, it limits the ability to make corrections to the data set. Along with positional accuracy, the collection of data in the field provides the most accurate method for checking completeness of a data set.

To provide high reliability, the targeted sample size was 1/3 the population up to a maximum sample size of 60. Samples were drawn randomly from the set of data entities. The sample selection process was affected by several characteristics. First, data samples were drawn from areas outside restricted areas ("live-fire" zone, impact areas, etc). Second, accessibility limited the ability to sample linear and area entities, so sample points along these entities were selected rather than attempting to measure the entire entity.

The exception to these restrictions was the water bodies data set. The boundaries of these entity objects were defined by the water level of the lakes or ponds at the time the data was developed. This level fluctuates depending on the time of year, rainfall amounts, etc. Any measurements taken in the field might be inconsistent with the source data due to these conditions rather than from a positional error. Therefore, water bodies were field checked only for completeness of the data set.

After sample selection, a point feature was created either at an intersection between linear representations of the same entity or between linear representations of different entities. Table 7 describes how each line and polygon data set was transformed into point data.

Table 7. Sample point generation.

Entity Type	Entity Type	Point Intersections
Training Area Boundaries	Polygon	Points were created where the installation boundaries cross with a major road.
Installation Boundaries	Polygon	A sample was conducted to choose 30 percent of the training areas. Points were then created where the sampled boundaries crossed a road.
Roads	Line	Points were created at road intersections.
Pipeline Crossings	Line	Points were created at the intersections between the pipe crossing and the actual pipeline.
Streams	Line	Points were created at stream confluences.

A relatively simple method for generating a random sample of objects was implemented. First, the point data was developed for those linear and polygon data sets that were to be collected as points in the field (as stated in Table 7). Next, any objects located within the "live-fire" zone and outside of the installation were removed from the source data. A random sample was then generated with the remaining records using a script within ArcView. Finally, x and y coordinates were added to the attribute table as a guide to find the relative location in the field and as a measure for horizontal positional accuracy conducted later in the report.

Collection of field data.

Field data collection was allotted 3 1/2 days. Five people split into two groups to subdivide the work. Field sheets with the attribute information from the sample sets, along with maps of each sample set and a Fort Hood ITAM Training Map were used in conjunction with two differential GPS units to help assist in pinpointing the true locations in the field. Once the locations were found, measurements were recorded with the GPS.

The two GPS units used for data collection were Pathfinder Pro XR and Pro XRS (Trimble Navigation Ltd., Sunnyvale, CA). The Pro XRS unit received real-time differential corrections from OmniSTAR satellite differential service while the Pro XR unit received its real-time differential corrections via a Frequency Modulation (FM) broadcast from the U.S. Coast Guard. All data were then post-processed in the lab to correct any inaccuracies with the real-time differential GPS positions using base station files from the Texas Department of Transportation in Austin. These corrections can be obtained by FTP at: ftp://ftp.dot.state.tx.us/pub/txdot-info/isd/gps. To collect positions as accurately as possible when in the field, field sessions were scheduled using Trimble's online mission planner. Data were collected when satellite availability and geometry were most favorable. Trimble's *General Reference Guide* (1996) explains that the Position Dilution of Precision (PDOP), a measure of how well the satellites line up in relation to one another, is acceptable between 0 – 8. As a set standard for the field data, entities were measured and collected only at times when the PDOP (satellite geometry) was equal to or below 7. The guide also states that the lower the satellites are positioned on the horizon, the lower the accuracy of the measurements being taken (Trimble 1996). The elevation mask, an elevation angle above the horizon that blocks out those satellites positioned below the set elevation angle, was therefore set to 15 degrees. Table 8 is a summary of the data collected in the field.

Table 8. Field collection summary.

Entity Type	# of records from the sample set	# of samples collected in the field
Stream Crossings	20	24
Low Water Crossings	13	16
Dams	65	27
Installation Boundaries	32	27
Training Area Boundaries	65	30
Pipeline Crossings	21	12
Teledrops	26	58
Roads	65	24
Stream Confluences	67	11

Discussion of results.

Table 8 does not represent a 1:1 ratio between the records in the sample set and the sample of entity instances collected in the field because certain complications arose while recording measurements in the field. No other attribute data or metadata was provided with the original test data sets, so it became difficult to associate the representation in the digital data with the actual entity in the field based on a spatial reference point alone. In some instances similar entities were within meters of each other and could not be affiliated with one particular entity object from the source data. Therefore, in many cases additional locations and entities were recorded to ensure a sufficient number of locations needed for the spatial accuracy testing.

It is also difficult to be certain whether the entity instances collected actually represent the data to be tested. Because formal definitions for entity types were lacking, measurements taken in the field may not actually relate to the correct entity instances. For example, no guidelines were provided for defining a pipeline crossing. While the test data only contained authorized crossings, in the field it was difficult to differentiate the authorized crossings from the unauthorized crossings (Figure 12). Many new crossings were found, but the physical appearances of the authorized and unauthorized crossings were identical. Again, many of the crossings were within meters of each other, so it was difficult to decide which authorized crossing in the digital test data matched with the relative location in the field.

Figure 12. Authorized vs. unauthorized pipeline crossings.

Furthermore, inaccessibility and unfamiliarity with Fort Hood made it difficult to collect certain entities within the available timeframe designated for field collection. Intermittent streams and relatively dense vegetation of some of the terrain prevented location of and access to many stream confluences. Rainfall during the survey period further limited access to certain terrain.

These particular obstacles were responsible for some less consistent and complete sample data sets than desired. Rural drop locations, stream crossings, and low water crossings are the most complete and highly representative field data sets of the entity types sampled. Other data sets have fewer samples than desired, thus affecting the confidence level of subsequent analysis.

Digitizing From Digital Media

Preparation.

When data collection in the field is limited due to environmental or weather conditions, time of day, equipment capabilities or inaccessibility caused by vegetation cover, land restrictions, etc., remote sensing interpretation and digitizing can provide a time efficient method for data collection. General sources for developing digital data include paper maps, aerial photos, digital orthophotography, satellite imagery, or digital elevation models. For the intended purposes of

this project, a source was needed that was of higher quality than the original source data set. The following sources were used to develop control data for the accuracy assessment examination:

1. 1997 color infrared (IR) Fort Hood IGAS digital orthophotography (2-ft resolution).
2. 1997 black and white Fort Hood IGAS digital orthophotography (2-ft resolution).
3. 1995 USGS 7.5-min Digital Orthophoto Quarter Quadrangles (DOQQs) (1-m resolution).
4. USGS 7.5-min Digital Raster Graphics (DRGs) (dates vary; 5-m resolution).

A different approach to sampling was implemented with the digitizing method for developing data. Instead of using a random sample across the entire installation, a selected set of areas was chosen. Entities within these areas would then be digitized to represent the control data for the accuracy testing. The selected areas were sampled from a grid that was developed by Fort Hood as a supplemental reference to 1997 digital orthophotography (Figure 13). The grid covers the entire extent of the installation and is divided into 113 cells. Each cell has an area of 4 sq mi. The following criteria were used to select the areas:

- randomly sampled cells
- sample set derived from 30 percent of the complete set of cells.

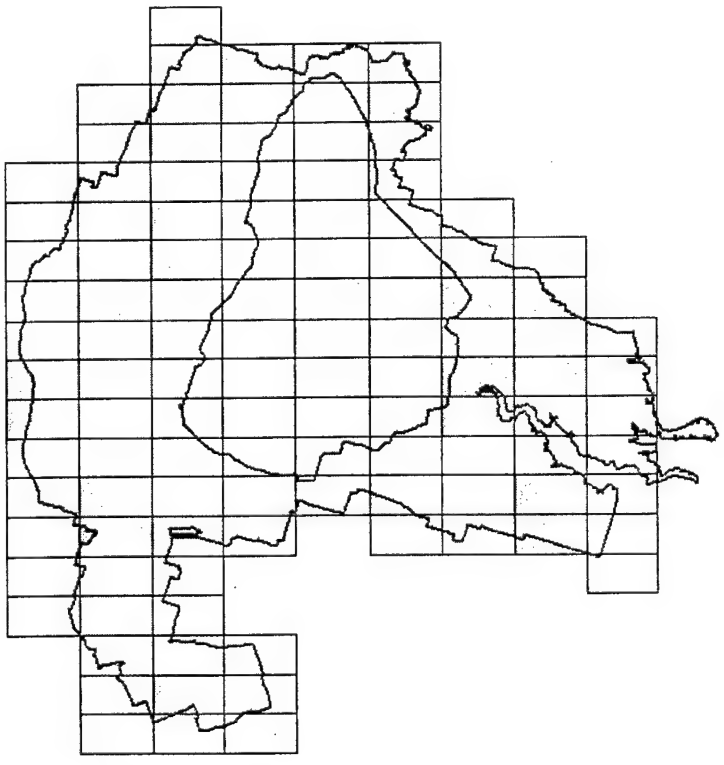


Figure 13. Sample grid.

All sampled grid cells needed to be more than 50 percent outside of the "live-fire" zone (restricted area) but more than 50 percent inside the installation boundaries. An ArcView script was used to generate a randomly sampled set of cells. The cells were removed after visual inspection if the criteria were not met.

A review of imagery for the data initially developed using this grid sampling technique found that streams were difficult to interpret out of context with the larger landscape. In some instances well-defined streams flowing through a grid were easy to recognize, but streams along the edges of the grid were often overlooked or misinterpreted. The grid boundaries disturbed the patterns of the network, and the discontinuous segments were difficult to identify. This digitized data could be used for accuracy checking, but confidence in the test would be reduced because there would be fewer matches between the control and test data.

An alternative sampling method was implemented for stream control data development from imagery. Sampling areas were defined and selected based on a small watershed data set provided by Fort Hood (Figure 14). Theoretically, examining a watershed area would help delineate the streams more accurately. The developer would be able to follow the stream from the headwaters down to the mouth of the watershed, improving the accuracy of the interpretation because both the detailed patterns of color, shape, and context and the overall pattern of the continuous network of streams could be seen. The following criteria were used to select the watersheds:

- randomly sampled areas
- sample set is derived from 35 percent of the watersheds (a larger sample was taken because, every time a sample set was generated, some watersheds were being selected along the outer boundaries that did not contain any streams).

An ArcView script was used to generate a randomly sampled set of watersheds. The watersheds were removed after visual inspection if the criteria were not met.

Data production.

A small staff of University of Illinois graduate and undergraduate students assisted in the digitizing of the control data. Aerial photography examples containing entities that were to be digitized were created as a training mechanism for the photo interpretation. Tutorials were also used to help familiarize the staff with software use.

Figure 14. Sample watersheds.

Five data sets to be used in the quality assessment testing were then digitized according to the grid or watershed (only for streams) sample area. Control data for two of the eight targeted entity types could not be created by this method of data collection. Rural drop locations were unrecognizable due to the resolution of the imagery. In regard to areas, what defined the installation or training area boundaries was unknown. Without this information, valid data sets could not be developed to accurately represent areas.

Once the sample areas were determined, a series of steps was taken to ensure that the data created were correctly interpreted:

1. For each data set generated, two individuals created the same data set independent from each other. No source data were used as an aid in interpretation of the images in order to avoid biasing the interpretation.
2. Once both individuals completed digitizing the entire sample set, a cross-comparison was performed. Any discrepancies were then reconciled into one final data set.
3. The reconciled data set was checked for any digitizing errors and topology was created. This data was then used as the representative control data set for a particular entity type in the quality assessment testing procedures.

Discussion of results.

Project staff developed the sample data sets over a period of 6 weeks. After the linear data sets were developed (i.e., streams and roads), point data sets were generated from the linear intersections (i.e., stream confluences and road intersections) as a secondary measure for testing positional accuracy. Table 9 summarizes the data created.

Table 9. Control data development.

Entity Type	Source	Sampling Method	Estimated Time Required To Develop Data
Pipe Crossings	1997 b/w DOQQ	Grid	1 day
	1997 color-IR DOQQ		½ day
Water Bodies	1997 b/w DOQQ	Grid	1 week
	1997 color-IR DOQQ		4 days
	USGS DRG		2 days
Streams	1997 b/w DOQQ	Grid	1 ½ weeks
	1997 color-IR DOQQ		1 week
	USGS DRG		2 days
	USGS DRG	Watershed	2 days
	1995 USGS DOQQ		1 ½ weeks
Roads	1997 b/w DOQQ	Grid	3 weeks
Stream & Low Water Crossings	1997 b/w DOQQ	Grid	1 week
	1997 color-IR DOQQ		5 days

The personnel assigned to develop the data were relatively unfamiliar with certain entity types, even with the set of examples produced as guides to help them distinguish key features on the aerial imagery. Even though stream networks can be identified quickly, the actual position and geometry of each stream channel within the network can be difficult to determine. This difficulty is due, in part, to the thick cover from the trees and surrounding vegetation that hides the actual channel and the lack of water flowing through these channels at certain times of the year (typical of hydrologic conditions in Texas). Stream crossings are defined by several different variables. Certain crossings may be misinterpreted or overlooked based on these three factors (Figure 15):

1. The ease of interpreting where the stream is positioned
2. The interpreter's definition of a tank trail
3. Determining a location where the tank trail and stream intersect.

Other entities are much more prominent and easier to distinguish on the imagery (i.e., roads, water bodies). These particular features may be under or over-represented, but the positions and boundaries of the features are much more clearly defined within the photographs.

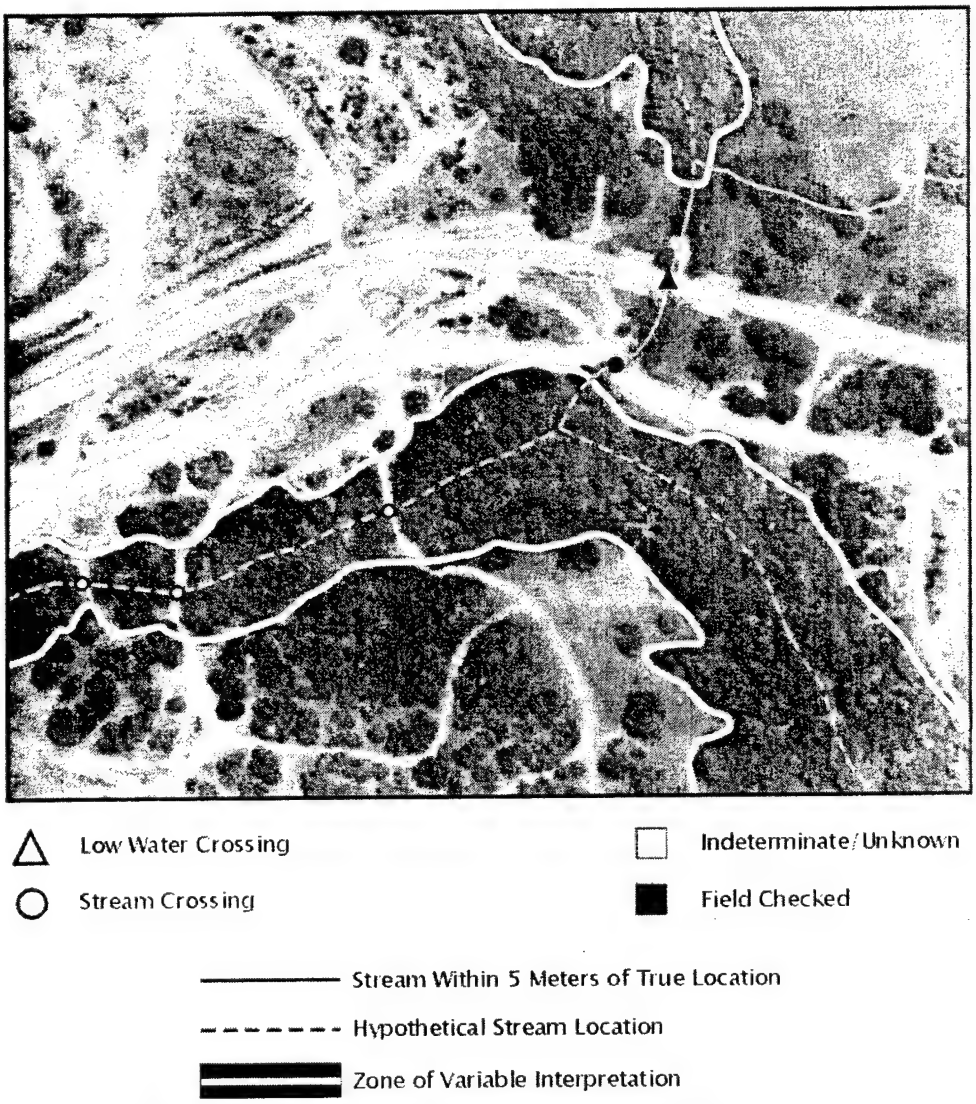


Figure 15. Difficulties in stream crossing identification for imagery.

Application of Accuracy Assessment Methods

Positional Accuracy

Table 10 lists each type of positional accuracy test completed on each of the eight entity types studied in this report. For area entities such as the installation and training area boundaries, an independent definition for boundaries did not exist, and ground locations could not be determined either in the field or from aerial photography.

Table 10. Positional accuracy tests completed by source.

	RMSE Test					*Line Buffer Test*					*Kappa Test*				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Roads	x					x									
Streams (by grid sample)	x	x		x		x	x		x						
Streams (by watershed sample)			x	x				x	x						
Pipe Crossings	x				x										
Stream/Low Water Crossings	x	x			x										
Teledrops					x										
Water Bodies											x	x		x	
Installation Boundaries					x										
Training Areas															

Source Reference Number: 1 = 1997 b/w Fort Hood IGAS digital orthophotography
2 = 1997 color-IR Fort Hood IGAS digital orthophotography
3 = 1995 color-IR USGS DOQQs
4 = USGS DRGs (various dates)
5 = GPS field-collected data

A detailed example for each type of positional accuracy test (as described in **Identify Control Data**, p 29) is demonstrated below.

Point test.

The following example is taken from the stream-crossing positional accuracy assessment using field-collected data.

Control data developed from the field survey was the only source used to assess the accuracy of the two Fort Hood stream-crossing data sets, strmxinging.shp and lowwaterxing.shp. Because these data sets were small, the two types were tested together. Several processes were required to prepare the test and control data for matching. All the preparation tasks were done using ArcView 3.2.

First the two Fort Hood data sets were merged. Each test object in the merged data set was given a unique identifier, and an attribute indicating its source data set. The Fort Hood data set had a total of 30 crossings, 13 designated "low water." Next the field-collected points were integrated into a single data set. Each field object was given attributes indicating its type, low water (prepared surface with a culvert) or stream crossing (no culvert), and a label indicating what object it represented from the Fort Hood data sets (Figure 16). Of a total of

40 field-collected crossings, 15 were designated low water. Finally, for each of the merged data sets, the coordinate locations for the objects were stored as attributes.

The objects from the merged test and control data sets were joined based on their identifiers and their stream crossing types. The join process resulted in 24 matched crossings, 10 designated as low water. The matched objects were selected, and the information critical for the RMSE calculation was exported to a dbase file:

Crossing identifier (crossing type and integer identifier)
Test data x coordinate
Test data y coordinate
Field data x coordinate
Field data y coordinate

The RMSE was determined in a standard spreadsheet that read the dbase file and applied the appropriate calculations. The spreadsheet for the combined stream crossing data sets is shown in Figure 17.

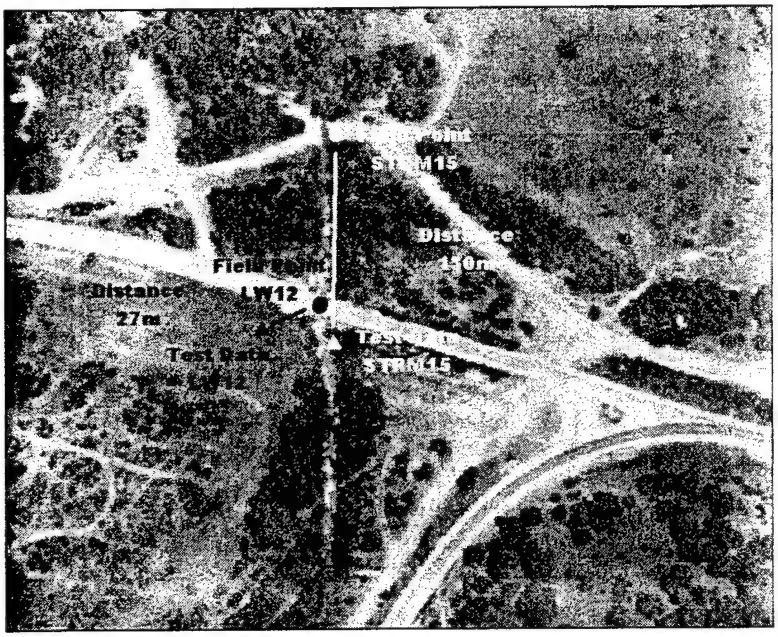


Figure 16. Sample point test data.

Horizontal Accuracy Statistic Worksheet for Ft Hood Crossings Data

Completed: 7/19/1999
By: NT

Enter Source/Field
Combined Dbase Filename: matches_nt.dbf

Enter column letters from combined dbase file:
Feature ID: C
Test x: A
Test y: B
Control x: D
Control y: E

Note: Control data is Field Collected Data from 5/99

# of Rows in dbase file: 0

Labels from dbase file: #REF! #REF! #REF! #REF! #REF!

C Point ID	D x (test)	E x (control)	F diff in x	G (diff in x) 2	H y (test)	I y (control)	J diff in y	K (diff in y) 2	L (diff in x) 2 + (diff in y) 2
#REF!	#REF!	#REF!	#REF!	#REF!	#REF!	#REF!	#REF!	#REF!	#REF!

Next, a series of buffers was developed around the control data set. The lengths of the buffers varied across sources and samples. For this particular source and sample, buffers with lengths from 0-40 m (in increments of 2) along with a 60-, 80-, and 100-m buffer were created.

After the buffers were generated, the original matched test data set was clipped by each buffer length. This helped determine how much of the test data was within a given distance of the control data. An attribute called "BUFFER" was then added to all objects within each of the 24 clipped data set's arc attribute tables (AATs). All objects within each data set were selected and assigned a value based on the buffer length to which they were clipped (as shown in the Figure 18 example for an object in the data set clipped by an 8-m buffer).

FNODE#	TNODE#	LPOLY#	RPOLY#	RIVCLIP8#	RIVCLIP8-ID#	LENGTH	BUFFER
20	22	1	3	3	91	559.012	8

Figure 18. Example from AAT table after clipping process is complete.

Once every object for every data set created had a buffer value assigned to it, all objects were appended into one data set. A statistical summary (using the STATISTICS command within *ARC*) was then generated to obtain the summed lengths of the objects based on their buffer value. This summary table was exported to a dbase file and brought into a standard spreadsheet (Figure 19) to calculate the percentage of the tested streams within the various buffer distances.

Kappa testing.

The following example is taken from the water body positional accuracy assessment according to the grid sampling method.

Control data developed from the 1997 Fort Hood IGAS color-IR digital orthophotography was one of three sources used to assess the positional accuracy of the Fort Hood data set alldams99bnd. Several processes were required to prepare the test and control data for matching. All the preparation tasks were done using Arc/Info 7.2.1 GIS software.

First any water bodies in the test data outside of the grid sampling areas were eliminated. Test and control data were next visually inspected to match the objects in each data set (by relative positional location). An aspatial match was not performed on the data due to the absence of attributes. Objects within the test data that did not have a match in the control data were removed. Likewise, any entities within the control data that did not have a match in the test data were removed. Only entities present in both data sets were saved.

Positional Accuracy Testing For Linear Features

Feature: Streams
Source: USGS 7.5 minute DRG's
Sampling Method: Watershed sample

Buffer Size	*Frequency*	*Sum of Lengths*	*% of Line Within Buffer*
0	378	268161.690766	100.00
2	4	114.615718	0.04
4	1201	47327.676870	17.65
6	1214	64820.976922	24.17
8	1227	80805.187532	30.13
10	1207	97192.324484	36.24
12	1191	112763.993813	42.05
14	1155	126707.492506	47.25
16	1122	140313.475707	52.32
18	1071	153986.080295	57.42
20	1008	166923.304629	62.25
22	978	177492.675907	66.19
24	912	188901.756244	70.44
26	868	198021.496041	73.84
28	834	205620.451366	76.68
30	776	213363.604876	79.57
32	746	219925.996425	82.01
34	693	226568.977396	84.49
36	651	231707.049227	86.41
38	606	236466.753069	88.18
40	579	240139.369723	89.55
60	431	258441.593583	96.38
80	386	262380.255792	97.84
100	380	263056.353341	98.10

Figure 19. Spreadsheet for testing positional accuracy with buffer/clip method.

The Kappa statistic was used to assess this polygon data. First, the entities needed to be combined and summarized by area into the four classifications (Table 2). To do this, attributes were assigned to the control and test data sets (TYPE1 and TYPE2, respectively). The value IN was assigned to all the objects in both the control and test data sets. The object defining the outer extents of the data set (the "world polygon") was assigned the value OUT. The two data sets were then combined (using the INTERSECT command within *ARC*), resulting in each object receiving one of the four combinations of classifications. A statistical operation (using the STATISTICS command within *ARC*) summed the areas together based on their classification. The summary table was then exported as a dbase file and brought into a standard spreadsheet. The area summaries were added to a matrix and formulas calculated the percentage of agreement between the control and test data, the "expected" fraction of agreement, and the Kappa statistic (Figure 20). (*NOTE*: the installation boundary was used as the outer extent. The P_{∞} [or P_{22} as shown in Figure 20] classification was derived by taking the installation area and subtracting it from the summation of the three other classifications' areas.)

Kappa Statistic Worksheet (Horizontal Accuracy Test for Polygons)

Directions: Insert values from dbase file into the yellowshaded cells.

Summary of Classification Values for IRDOQ data

Kappa	Frequency	Area
P_{22}	1	883760698.9
P_{21}	172	547484.8203
P_{12}	95	12928.73438
P_{11}	59	583484.875

Note : The actual P_{22} is equal to the area of the installation minus the water bodies.

Therefore, P_{22} = 883760698.88231

Probability Matrix for Fractional Amounts

		Classified by second source		
		1	2	
Classified by first source	1	0.000659376	1.46103E-05	0.000673986
	2	0.000618694	0.99870732	0.999326014
		0.00127807	0.99872193	1

Final Results

$P_e = (P_{1+})(P_{+1}) + (P_{2+})(P_{+2})$

$P_e = 0.998049667$

$P_o = P_{11} + P_{22}$

$P_o = 0.999366696$

Kappa = $(P_o - P_e) / (1 - P_e)$

Kappa = 0.675284236

Figure 20. Spreadsheet for testing positional accuracy with Kappa.

Attribute Accuracy

Performance of attribute accuracy assessments was limited by the lack of explicit attributes. A number of attributes were implied by the separation of entities between several data sets. However, the ability to assess these attributes was limited by whether they could be sufficiently defined and either observed in the field or on imagery or verified from an independent source. Of the three implied attributes (stream crossing type, stream type, road type), only road type was sufficiently defined and observable in the field.

A road type attribute existed for the major roads within the Fort Hood installation boundary. A complete set of major roads, *roadmajall.shp*, contained primary and secondary, nonresidential vehicle transportation routes for Fort Hood and related quadrangles. The road type attribute for these objects was implied in their separation between two additional data sets: *roads_imp.shp* and *roadssec.shp*. A definition of the attribute classes was developed based on inspection of the data sets, review of the SDS standard for roads, and a visual inspection of the roads (Table 11).

Table 11. Attribute classes for roads.

Road Type	Data Set	Definition
Primary	Roads_imp	Main vehicle transportation routes, paved
Secondary	Roadssec	Nonresidential secondary vehicle transportation routes, gravel graded

A third classification type, "other," needed to be defined for the assessment (see Figure 21). This classification did not represent a "true" road type, but was used as a grouping mechanism to summarize inconsistencies in objects represented across the three data sets. An object in the test data was considered type "other" if (1) it existed in both the improved and the secondary data set (classified twice) or (2) it existed in the major roads data set but did not exist in either the improved or secondary data sets. If the road type characterization had been implemented using an attribute rather than with separate data sets, this double classification would not have occurred and uncharacterized objects would have been easy to identify.

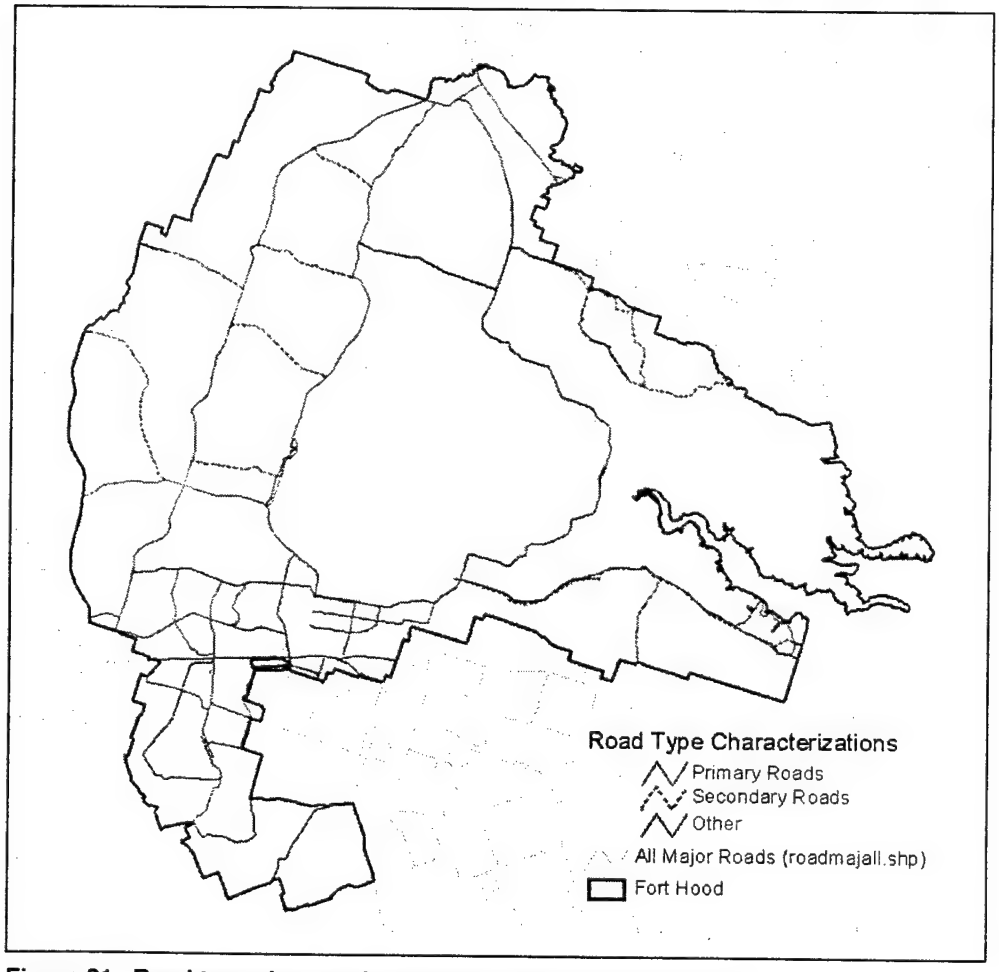


Figure 21. Road type characterizations from Fort Hood data sets.

The attribute characterizations were verified during field data collection. A target control data set of roads was not predetermined with random sampling. Rather, road type was recorded for any major road segment traversed in the course of field collection of other data. Field collection teams each had a large format map with the major roads symbolized by classification type per the test data.

Determination of the attribute accuracy was complicated by the fact that the data sets describing major roads were topologically inconsistent. The field survey inspected approximately 40 road segments. The data sets, however, contained 330 objects to define these 40 segments. To correctly calculate a meaningful attribute accuracy measure, the field survey data needed to be realigned with the existing data sets. Once this was accomplished, attribute accuracy could be quantified using either the percent correctly classified or the Kappa statistic (Table 12).

Table 12. Results of attribute accuracy assessment for road characterizations.

Road Type	# of Objects Observed in Field	# of Objects Classified in the Data Sets	# of Objects Classified Correctly in the Data Sets
Primary	265	195	194
Secondary	65	87	53
Other		48	
Total	330	330	247
Percent Correctly Classified (247/330):			75%

		Classified by *Test* data			
		Primary	**Secondary**	**Other**	
Classified by *Control* data	**Primary**	0.588 (194)	0.103 (34)	0.112 (37)	0.803
	Secondary	0.003 (1)	0.161 (53)	0.033 (11)	0.197
	Other	0.000 (0)	0.000 (0)	0.000 (0)	0.00
		0.591	0.264	0.145	370
Percent in agreement (0.588+0.161+0.000)					0.749
Percent expected agreement (0.803*0.591+0.197*0.264)					0.526
Kappa statistic ((PA – PE) / (1 – PE)					0.470

Logical Consistency

Arc/Info 7.2.1 GIS software was used to check the logical consistency of the data sets being reviewed. Unlike ArcView, Arc/Info allows the assessor to examine and correct any topological errors that may have occurred. Data was checked for duplicate lines representing the same entity or sliver polygons that are sometimes created when duplicated lines have not been removed. Two types of nodes,

dangles and pseudos, were examined for errors. Dangles, created when digitized linear objects stop short of, or extend past, an intended intersection point, were corrected or removed from the data. Those pseudo nodes that did not represent island polygons were eliminated. Labels were checked with the LABELERRORS command to be certain that there were not any polygon objects with two labels or no label at all. Also, any linear segments that did not appear to be part of an object or were not put there intentionally (usually a by-product of a low/high snap tolerance during data creation) were again corrected. Finally, topology was created using the CLEAN command within Arc/Info in preparation for final delivery.

Attribute consistency was minimally assessed for certain data sets. Because Fort Hood created a new data set for each particular classification of an entity, attribute accuracy was difficult to assess. Overlaps in classification were found when data sets representing different classifications of the same entity were examined together. For example, the principal stream data set used by Fort Hood is *riverall.shp*. This data set did not have any attributes distinguishing stream types within the data. Other data sets were provided by Fort Hood that seemed to be derived from *riverall.shp* with some minor topological additions. Each of these data sets represented a different classification of streams, such as small rivers and large rivers. Independent examination of each data set revealed no problems with attribute consistency because no attributes were present. However, when the data sets were joined together, some objects appeared twice and were labeled as both small and large. A 10.23 percent inconsistency was caused by this overlap of large and small streams classification.

When data are being created for an entity, it is recommended that all classifications be represented with attributes and stored within one data set. This practice guards against inconsistencies that occur when an entity type is broken into multiple data sets for classification purposes.

Completeness

Completeness generally can be assessed only when there is sufficient information about a data set's lineage, currentness, and entity type definition. Most data sets provided by Fort Hood lacked this critical information, which enables the association of objects in the test data to objects in the control data. Currentness was particularly a problem. Without knowledge of the Fort Hood data set's currentness, matching it to an older or newer control data set would result in an erroneous assessment of completeness. Another issue was entity type definition. For example, field collection of rural drop locations found multiple drop lines at a

single site. With no attributes or data set description, it was impossible to determine whether the test object represented all four of these lines or only one.

In one data set, the water bodies data (*alldams99bnd.shp*) did not need an explicit definition because the entities could be so clearly identified from imagery. A source history was developed by examining different years of imagery and available historic data sets (e.g., *alldams97.shp*). Each object in the data set was assigned a date of when it was added to the data set. A completeness assessment was then performed by comparing the test data set to the control data set produced from 1997 imagery. A total of 79 water bodies were identified on the 1997 Fort Hood IGAS color-IR digital orthophotography for the control data set. The test data set contained a total of 96 water bodies within the same sample area. Of these 96 records, 14 were removed from the completeness assessment because they were collected from GPS locations after the 1997 imagery was photographed; thus, they might not have appeared on the imagery used to create the control data. Including them would have increased the overstatement by 10 percent and lowered the completeness of the data by 13 percent. The remaining 82 objects were matched (by relative spatial location) to the control data set. A total of 58 matches occurred between both data sets. From this, a percent of omission and commission was derived:

PO = 73.42% completed

PE = 29.27% overstatement

Because no other independent source was available to describe what attribute types are required for the water bodies data, the SDS was used as a guide to determine which attributes types should be included within the data. Seven attributes were recommended by SDS including name, description, type, permanence, relative up and down mile markers, and the year developed (if man-made). Of these seven types, two types matched with the test data: name and description. Therefore by the SDS standards, the attribute completion for water bodies is:

PC = 28.6% completed.

5 Results

Chapter 4 described the methods used to document the qualities of data with respect to positional accuracy, attribute accuracy, completeness, and consistency. Examples of calculations for these methods were provided. Table 13 summarizes the assessments for the selected data sets. To ensure the sharing of the information obtained in performing the QA/QC procedures, metadata records were created for each of the selected Fort Hood data sets. The metadata records reflect the inspection process and the application of the accuracy assessments described in Chapter 4. Metadata was created using CORPSMET95, a software tool for generating FGDC-compliant metadata files.

Table 13. Assessment results for selected Fort Hood data sets.

				Point			Line		Polygon
Data Category	Data Set	Sampling Method	Control Data Sources	# points	RMSE (meters)	NSSDA (meters)	Epsilon buffer size in meters	Epsilon % of line within buffer	Kappa
Installation Boundary	universe.shp		field	9	15.82	27.38			
Training Area Boundaries	trnaful_poly.shp		could not use	-	-	-			
Roads	Roadmajall.shp	grid	imagery				32	94.93	
							34	95.18	
Roads	road.shp	grid	imagery	405	30.66	53.07	100	94.70	
Surface Hydrology	riverall.shp	grid	DOQ imagery	174	44.58	77.15	60	93.06	
							80	95.33	
		watershed	DOQ imagery	144	46.27	80.09	100	94.53	
		grid	DRG imagery	122	37.94	65.67	40	88.22	
							60	96.51	
		watershed	DRG imagery	99	42.12	72.9	40	89.55	
							60	96.38	
		grid	IRDO imagery	62	54.98	95.16	80	94.39	
							100	96.14	
Dams	alldam99bnd.shp		DOQ imagery						82.00%
			DRG imagery						77.00%
			IRDO imagery						68.00%
Stream Crossings	strmxing.shp/ low-waterxing.shp		field	24	65.83	113.95			
	strmxing.shp		field	14	74.08	128.21			
	Lowwaterxing.shp		field	10	52.16	90.27			
Pipeline crossings	pipexing.shp	entire pipe-line	imagery	9	50.21	86.9			
Rural Drops Locations	teledrop.shp		field	11	91.08	157.65			

6 Conclusions

Summary

The primary goal of this research was to develop and test methodology to assess, report, and improve the quality of spatial data used in Army installation ITAM databases. The specific tasks include: identification and performance of QA/QC procedures on Fort Hood ITAM GIS data layers; documentation of core ITAM GIS data layers using the FGDC Content Standard for Digital Geospatial Metadata.

The approach used a process of assessing the status and quality of selected existing data sets based on current standards and the research literature, investigating methods and resource requirements to improve the quality of these data sets, and documentation of the findings. Results of the assessment and improvements are reported according to the FGDC Content Standard for Digital Geospatial Metadata using the CORPSMET95 software and other documentation. To initiate this effort and provide a framework for future efforts, the project developed and tested procedures for performing a post-development assessment of the quality of selected data from the ITAM data set at Fort Hood, Texas. The methodology for QA/QC, presented above, incorporates practical and accepted approaches promoted in various standards (such as the FIPS Spatial Data Transfer Standard) and in current research literature.

Conclusions

If the research hypothesis were "Is it possible to assess the quality of existing spatial data sets with a high degree of reliability?" the initial results from this study would have to lead to the ambiguous response, "it depends." Adequate documentation of data is a systemic problem in all areas of information systems; the geographic information community, including the Army installation geographic information communities, are not likely to be different. The data sets surveyed in this study suffered from a lack of documentation on the contents and history of the data stored in the files. While spatial data is perhaps minimally self-documenting because of the graphic representation of data, this method is by no means sufficient. The study results indicate that, in general, documentation

of data is inadequate to understand the nature, and thus the utility, of the ITAM geospatial data sets.

Reasons for insufficient documentation are numerous and can include ambiguous lines of responsibility for maintaining data records, competing demands for human resources, insufficient training, and numerous other factors tied to a particular organization. Added to the equation is the lack of comprehensive standards for accuracy relative to a specified purpose, lack of accepted (or adopted) standards for organization and classification of data, and lack of available methods and tools for assessing and documenting data quality factors.

The ability of the Army to adequately address future needs for the management of its land resources is directly tied to its ability to maintain adequate information resources. Without intelligence, there can be no effective planning or even reaction to events that take place. Ensuring the quality of its information resources is fundamental to any meaningful analysis the Army and its contractors might wish to carry out.

Recommendations

It is risky to base recommendations on the limited exposure to ITAM's GIS data sets. The most obvious recommendations demand more study. Nonetheless, given the breadth of exposure to installation databases, it is relatively safe to say that Fort Hood is a typical installation. Below is a summary of recommendations for further research that is focused on improving the ability of an installation to maintain its geographic data sets.

1. Expand the inspection of data sets to determine the status of QA/QC procedures and the current state of data sets.
2. Explore approaches to organization of data sets to facilitate the conduct of QA/QC procedures, especially that associated with lineage.
3. Explore alternative approaches to data development given changes in data sources and methods (e.g., develop hydrology networks from high-resolution digital elevation models rather than interpretation of remote sensing information).
4. Explore application of or develop tools for QA/QC assessment to be embedded within COTS software products adopted for use by the Army.
5. Investigate and develop approaches to integrating metadata with the data (e.g., "smart-data" – data that are self-describing to applications).

With the Army's investment in and potential dependence on its geospatial data, the need for adequate maintenance of its information infrastructure is apparent.

This study demonstrated that, while the documentation of data is highly variable, and generally less than promulgated standards require or suggest, it is possible to conduct reviews of legacy datasets to reconstruct elements of documentation and assessment of quality for subsequent use. It is hoped that this research will have two outcomes for the Army. First, it will identify directions for development of valid and usable methods of assessment to recapture the value of the Army's investment in existing datasets. Second, it will encourage data managers to provide complete documentation of data at the time it is developed, in part by providing readily useable guidance through the various components of documentation and QA procedures.

Data represent a substantial part of the Army's infrastructure. In the same manner that one would not build a building without documentation, or maintain financial records without procedures for audit and control, maximization of the utility of information resources requires investment in data documentation.

References

Brassel, K., F. Butcher, E.M. Stephan, and A. Vekovski, "Completeness," in *Elements of Spatial Data Quality*, edited by S.C. Guptill and J.L. Morrison (New York: Elsevier Science Ltd, (1995), pp 81-108.

Caspary, W. and R. Scheuring, "Positional Accuracy in Spatial Databases," *Comput. Environ. and Urban Systems,* Vol 17 (1993), pp 103-110.

Congalton, R.G., "A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Maps Generated from Remotely Sensed Data," *Photogrammetric Engineering and Remote Sensing,* Vol 54, No. 5 (1988), pp 593-600.

Congalton, R.G., "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data," *Remote Sens. Environ.*, Vol 37 (1991), pp 35-46.

Federal Geographic Data Committee (FGDC), *Geospatial Positioning Accuracy Standards; Part 3: National Standard for Spatial Data Accuracy,* FGDC-STD-007.3 (FGDC, Washington, DC, 1998). http://www.fgdc.gov/standards/status/sub1_3.html

FGDC, *Content Standards for Digital Geospatial Metadata (version 2.0)* FGDC-STD-001 (FGDC, Washington, DC, 1998). http://www.fgdc.gov/standards/status/csdgmovr.html

Finn, J.T., "Use of the Average Mutual Information Index in Evaluating Classification Error and Consistency," *International Journal of Geographical Information Systems,* Vol 7, No. 4 (1993), pp 349-366.

Gaertner, P.J., "Quality Assurance Procedures Vital to Conversion Projects," *GIS World,* Vol 6, No. 9 (1993), pp 34-37.

Goodchild, M.F., "Attribute Accuracy," in *Elements of Spatial Data Quality*, edited by S.C. Guptill and J.L. Morrison (New York: Elsevier Science Ltd., 1995), pp 59-80.

Goodchild, M.F. and G.J. Hunter, "A Simple Positional Accuracy Measure for Linear Features," *International Journal of Geographical Information Science,* Vol 11, No. 3 (1997), pp 299-306.

Greenland, A., R.M. Socher, and M.R. Thompson, "Statistical Evaluation of Accuracy for Digital Cartographic Data Bases," in *Proc. Auto-Carto 7: Digital Representations of Spatial Knowledge* (Washington, DC, 1985), pp 212-221.

Guptill, S.C., "Temporal Consistency," in *Elements of Spatial Data Quality*, edited by S.C. Guptill and J.L. Morrison (New York: Elsevier Science Ltd, 1995), pp 153-166.

Kainz, W., "Logical Consistency," in *Elements of Spatial Data Quality*, edited by S.C. Guptill and J.L. Morrison (New York: Elsevier Science Ltd, 1995), pp 109-138.

Leung, Y. and J. Yan, "A Locational Error Model for Spatial Features," *International Journal of Geographical Information Science,* Vol 12, No. 6 (1998), pp 607-620.

Salgé, F., "Semantic Accuracy," in *Elements of Spatial Data Quality*, edited by S.C. Guptill and J.L. Morrison (New York: Elsevier Science Ltd, 1995), pp 139-152.

Skidmore, A.L. and B.J. Turner, "Map Accuracy Assessment Using Line Intersect Sampling," *Photogrammetric Engineering and Remote Sensing,* Vol 58 (1992), pp 1453-1457.

Spatial Data Accuracy Handbook - Draft: Implementing the National Standard for Spatial Data Accuracy, Minnesota Governor's Council of Geographic Information, October 9, 1998.

Spatial Data Standard, CADD/GIS Technology Center, U.S. Army Corps of Engineers, Engineer Research and Development Center, http://tsc.wes.army.mil/.

Spatial Data Transfer Standard, ANSI-NCITS-320-1998, American National Standards Institute, http://www.ansi.org.

Stanislawski, L.V., B.A. Dewitt, and R.L. Shrestha, "Estimating Positional Accuracy of Data Layers Within a GIS Through Error Propagation," *Photogrammetric Engineering & Remote Sensing,* Vol 62, No. 4 (1996), pp 429-433.

Trimble Navigation Ltd, *General Reference Guide* (1996).

Tveite, H. and S. Langaas, "An Accuracy Assessment Method for Geographical Line Data Sets Based on Buffering," *International Journal of Geographical Information Science,* Vol 13, No. 1 (1999), pp 27-47.

U.S. Department of Interior, "A Study of Land Information," prepared by U.S. Department of Interior in accordance with Public Law 100-409, November 1990.

Glossary

AAT	Arc attribute table
AMI	Average Mutual Information
ANSI	American National Standards Institute
BOS	Buffer and overlay statistic
CAD	Computer aided design
COTS	Commercial off-the-shelf (software)
CSDGM	Content Standard for Digital Geospatial Metadata
DLG	Digital Line Graphics
DOQQ	Digital Orthophoto Quarter Quadrangle
DRG	Digital Raster Graphic
EO	Executive Order
ESRI	Environmental Systems Research Institute
FGDC	Federal Geographic Data Committee
FIPS	Federal Information Processing Standard
GIS	Geographic Information System
GMS	Geographic Modeling System
GPS	Global Positioning Satellite
IR	Infrared
ITAM	Integrated Training Area Management
Kappa	A calculation of percent correctness of a map; allows for comparisons with other maps.
NCDCDS	National Committee for Digital Cartographic Data Standards
NSDI	National Spatial Data Infrastructure
NSSDA	National Standard for Spatial Data Accuracy
PA	Percent of agreement
PCC	Percent correctly classified
PDOP	Position Dilution of Precision
PE	Percent of excess

QA	Quality assurance
QC	Quality control
RMSE	Root mean square error
SDS	Spatial Data Standard
SDTS	Spatial Data Transfer Standard
TIGER	Topologically Integrated Geographic Encoding and Referencing System
USGS	United States Geological Survey

Appendix A: Processes of Control Data Creation

Control data can be acquired through a variety of sources depending on the budget available and the degree of accuracy necessary. The two most common ways of acquiring control data are to obtain data that has already been developed from another source, or develop new data sets.

Control data from secondary sources can be an efficient way to save money and time. However, certain precautions should be taken when using secondary data as a control set:

1. It must have at least one of the five characteristics that make it a source of higher quality as stated in the section on identifying control data (p 28).

2. It must have the same entity type/representation of the test data.

3. In some instances, data already developed and readily available may represent only a portion of the study area, or represent a portion of the feature types in the test data. The control data needs to have enough representative features to match the sampling requirements for statistical reliability.

4. A number of secondary data sets may not have metadata. Without knowledge or documentation of the data's history or accuracy, it cannot be an assured representative of a higher quality data set.

When readily available data (meeting the criterion specified above) cannot be located, control data sets can be developed to suit the needs of the assessor. Data sources can include field collection or digitizing from various media. Different methods of collecting data can be implemented depending on time constraints, spatial extent, and ease of collection for a particular entity type.

The collection of data in the field provides the most accurate method of data development. The most efficient method of collecting data quickly and accurately is the GPS (Global Positioning Satellite) receiver. Before data collection begins, a series of preliminary tests should be run on a sample set of features. This

establishes the difficulty of identifying and capturing the entity type, and optimizes the time and accuracy of collecting the data when in the field.

Various sampling methods are available to assess data:

1. Select a random sample of entity objects based on the test data set to verify in the field. This method can provide a good representation of features across large areas, but it requires a lot of effort due to the distribution of features. Also this method does not allow for completeness checking because the samples chosen were derived from the test data. The only type of completeness checking that can be determined is whether or not the features within the test data actually exist in the field. This method is recommended for smaller study areas to cut down on the amount of time it takes to travel from feature to feature.

2. Select a random sample of areas (cluster sampling) and collect all entity objects located within these selected areas. This method provides a more efficient and time conservative method of data collection but, because it is less evenly distributed than random sampling, spatial bias may not be identified in the data.

3. Select a sample from stratified groups of test data. The stratified groups represent subclassifications based on knowledge of differences between entity objects. Differences might be attributed to spatial distribution, feature type, source variability, or ease of collection or identification. Random samples can be selected from each group and independently tested. Rather than a random sample, the sample can be weighted to reflect the expected importance of each group. Alternatively, a stratified sampling method is recommended for data sets that might contain different errors for these different classes. Knowledge of the data set is key to applying this approach and improving the quality assessment.

4. Once a sampling methodology is determined, develop a data dictionary for the GPS for each entity type to be sampled in the field. A data dictionary assists in documenting attribute information about a particular feature quickly while the GPS is simultaneously collecting positions. Any attributes being tested that can be collected or verified in the field should be placed into the data dictionary.

Set aside 1 or 2 days to perform field tests. Take this time to become familiar with the study area, including the terrain, spatial extent, and accessibility of each entity type being tested. Then determine an efficient way to collect each

entity (point, line, or polygon) if it is at all possible to collect. Some features, such as streams, may be impractical to collect in the field. In some cases, alternative methods for testing linear and polygon features can be implemented to test positional accuracy. Point features can be generated from well-defined points on a line, such as a stream confluence or a road intersection, to determine the positional accuracy of the linear feature.

Field testing also includes checking the reliability and efficiency of the data dictionary for the GPS. Testing should be done to make sure the data dictionary coincides with what is actually being portrayed in the field. This time is also appropriate to clarify any questions one may have about the definitions of any entity type. If it is discovered after the fact that the features that were collected in the field misrepresent the data being tested, then the control data may be useless for applications toward any accuracy assessment.

Updates should then be made to the data dictionary to reflect any field test findings. These updates would include any changes made to collect an entity by an alternative feature type or any changes to certain characteristics that could be added or deleted from the feature description section of the data dictionary. Field sessions should then be planned that are based around times when the availability and positioning of the satellites provides the most accurate GPS readings. Trimble provides an online mission planner at http://www.trimble.com/.

After changes are made and a plan is framed for collecting and documenting the entities in the field, the samples are ready to be field collected for each entity. Field sheets (containing attribute information), maps of the sample sets for each entity, aerial photography, and road and topographic maps of the study area are recommended as supplemental aids in locating features in the field.

Once the GPS data is corrected (real-time, post-processed, or both) and downloaded into the appropriate format, it is prepared for testing purposes.

Data collection in the field can be limited due to environmental or weather conditions, time of day, equipment capabilities or inaccessibility caused by vegetation cover, land restrictions, etc. Remote sensing data interpretation and digitizing can provide a time efficient alternative method for data collection. Sources for developing digital data include paper maps, aerial photos, digital orthophotography, satellite imagery, or digital elevation models. Again, it is important to note that the source chosen to develop the data must be a source of higher quality. If the quality of the test data is unknown, it is usually best to use the most recent version of digital orthophotography available.

A sample method needs to be determined using the same principles as field collection. The clustering or stratified sampling methods are recommended (in certain cases) when collecting data by digitizing. This allows the developer to scan through selected areas and capture all features of a particular entity regardless of whether they are in the test data or not. An accurate assessment of completeness can then be established and features can be corrected in the test data if desired. This also provides a reliable set of features for checking the logical consistency of linear networks. A variety of boundaries defining the areas can be implemented when selecting areas for the sample. For example, grids that are geometrically equal can be used for certain features such as water bodies, land use divisions may be used to develop a roads data set, or watershed boundaries may be used in developing control data for streams.

The personnel involved in developing the data should then be trained in such areas as aerial photo interpretation or the software used for creating the data. Definitions of all entities should be clearly presented and illustrated to remove any miscalculations in position or inhabitance. Some data sets may not be able to be developed through digitization methods. If imagery is used as a source, certain entities (mainly point feature types) may not be able to be distinguished due to resolution and the spatial extent of the feature. Other features that may not be able to be identified are those features concealed by the canopy cover from trees and surrounding vegetation. For optimal accuracy and identification, it is recommended that infrared color imagery be used when features are difficult to ascertain.

Once the sample sets for each entity are chosen, establish a set of guidelines to be used by the digitizers during the creation and editing of the. Steps should include checking for any digitizing errors produced during creation of the data (i.e., dangles, unclosed polygons, unwanted pseudo nodes, etc.) followed by the creation of topology for spatial analysis. A redundancy check of all newly created data sets is highly recommended as a precautionary measure to ensure the features represented in the control data truly match those defined in the test data. This task should be completed by another individual, preferably someone with more experience in aerial photo interpretation and GIS software tools. Updates should then be performed before the finalized version is used for the accuracy testing procedures.

Appendix B: Detailed Inspection Report

Inspection of Spatial Data

Preparing for "QA/QC Procedures on Fort Hood ITAM GIS Data Layers"

Prepared for:

US Army Construction Engineering Research Lab Champaign, IL under
USACERL Contract No. DACA88-97-D-004,
Task Order 0009: *QA/QC Procedures on Fort Hood ITAM GIS Data Layers*
Kelly Dilks, POC

Prepared by:

Geographic Modeling Systems Lab
University of Illinois at Urbana-Champaign, Urbana, IL
Douglas M. Johnston, PI Diane M. Timlin

22 April 1999

Overview

Data inspection is an important preparation task for this project. Information gained will enable the development and testing of methodology to assess, report and improve the quality of spatial data used in Army installation Integrated Training Area Management (ITAM) data bases.

The data set selected for analysis is from Fort Hood, Texas. Fort Hood's spatial data is typical of much spatial data: data sets were developed independently based on a particular need at a particular time, and metadata is scarce. The current users of this data have limited knowledge of the development methods and any related quality assurance or control mechanisms. A detailed inspection of the actual spatial data is intended to:

- help determine lineage
- identify gaps in the existing data
- clarify relationships between data sets
- assist in the selection of appropriate data sets for QA/QC testing

Spatial data for the Fort Hood was received in two transfers in February 1999. The first was a compact disc (CD) referencing over 600 MB of various geographic data. Unfortunately an error in the CD meant only about 50% of the data was accessible. A second set of data was delivered through an electronic transfer (File Transfer Protocol [FTP]). This data consisted of 37 data sets in ESRI's ArcView product's proprietary format, called "shape" files.

The data inspection focused on the 37 files received in the electronic transferred, as this data was specifically selected from the substantial library of data sets at Fort Hood. Files from the initial transfer were added to the inspection if they seemed particularly relevant or contained data not represented in the primary data sets.

The information presented in this document is the result of a detailed data inspection using functions available in ArcView. Topological comparisons across data sets were performed with spatial joins; attribute comparisons across data sets were performed with table joins and queries.

Data sets examined in this review have been organized according to the prioritized list of data sets to be addressed by the project:

1. installation boundary
2. training area boundary
3. surface hydrology

4. stream & pipeline crossings
5. rural drop locations
6. roads.

Installation Boundary Data (See Map Layout #1*)

Three applicable data sets in the FTP transfer:

Universe.shp	50 polyline features
Boundary.shp	31 polyline features
1954prop.shp	2203 polyline features

A fourth file was delivered in a later transfer:

Universe.shp	1 polygon feature

The originally delivered *Universe* contains polyline features defining a single area to represent the bounds of the installation. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (typical topological field initialized by ARC/INFO; ranges from 1 to 50)
tnode_ (typical topological field initialized by ARC/INFO; ranges from 1 to 50)
lpoly_ (typical topological field for polygon data initialized by ARC/INFO; all set to 0)
rpoly_ (typical topological field for polygon data initialized by ARC/INFO; all set to 0)
length (ranges from 1.275 to 22903.282; measure unknown)
universe_ (typical id fields initialized by ARC/INFO; ranges from 1 to 50)
universe_i (typical id fields initialized by ARC/INFO; ranges from 1 to 49 – 2 lines are labeled 1)

* Map layouts are shown at the end of each section.

Similarly, *boundary* contains polyline features defining a single area to represent the bounds of the installation. Attributes in the data table indicate the data came from an ARC/INFO source, but most of the fields are 0 value:

fnode_, tnode_, lpoly_, rpoly_ (all set to 0)
length (ranges from 188.057 to 22903.281; measure unknown)
boundary_ (ranges from 1 to 31)
boundary_i (ranges from 1 to 31)
data (all set to blank)

A spatial comparison between *universe* and *boundary* found matches for 17 features. Attributes matched only on the length field.

1954prop contains polyline features defining many small areas within and adjacent to the installation bounds defined by *boundary*. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 1 to 1642)
tnode_ (ranges from 1 to 1642)
lpoly_, rpoly_ (all set to 0)
length (ranges from 1.048 to 2763.195; measure unknown)
z954prop_ (ranges from 1 to 2203)
z954prop_i (ranges from 1 to 2752)

The replacement *Universe* contains a single polygon feature defining the extents of the installation. The data was likely produced from an ARC/INFO polyline cover:

id (2)
area (886636159.706)
perimeter (264714.497)
hectares (88663.616)

On delivery, this file was described as follows:

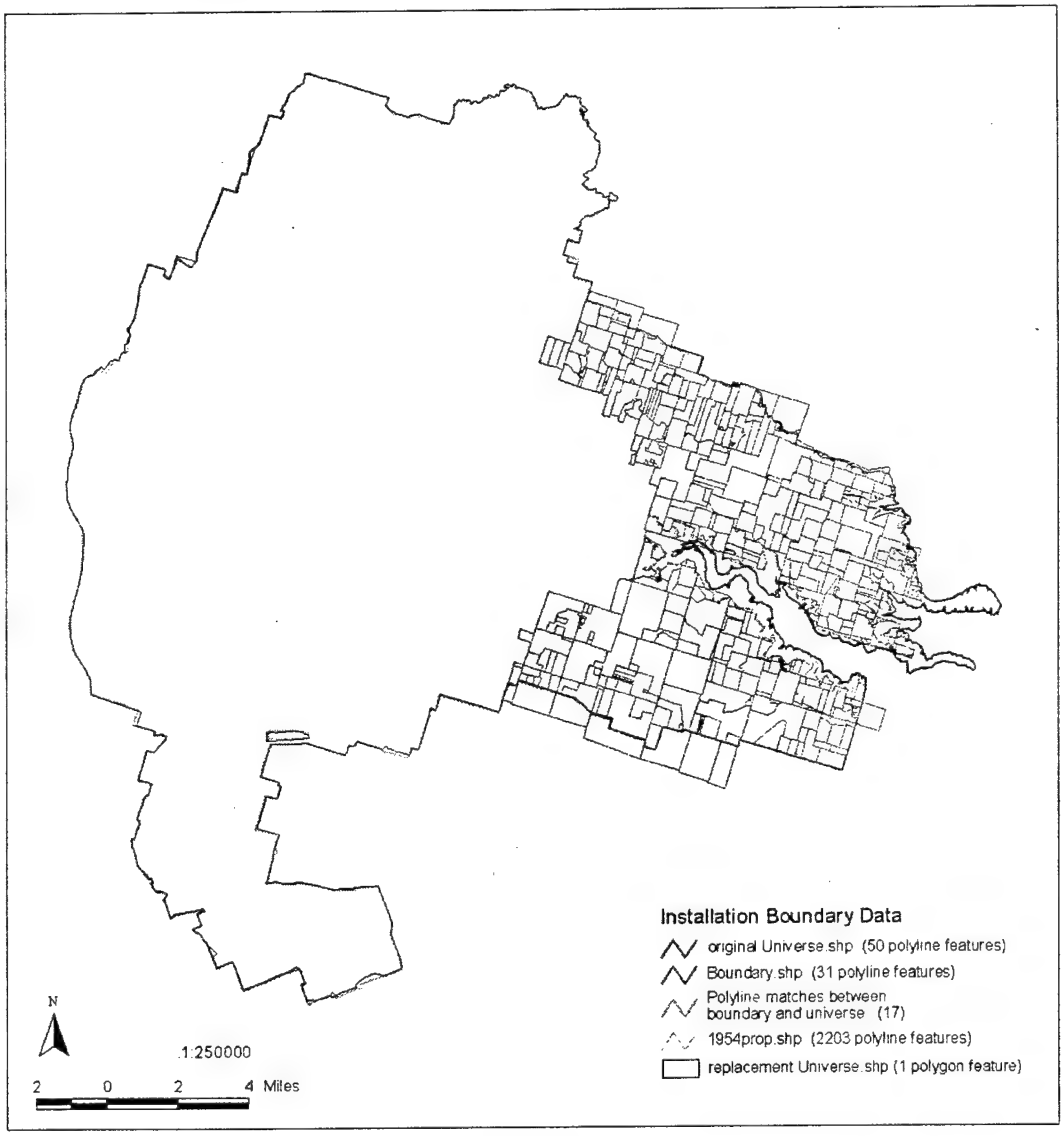
> Includes all land as a line file and has been updated with Aug 1997 2.5m DOQQ's (heads-up digitized – center line roads and stream channels)- includes Corps of Engineer property and Fort Hood property.

Without access to the original line data, spatial comparisons between this and other data cannot be performed. However, a visual comparison shows that the

general extent and geometry of the polygon is similar to the original *universe* data set. This file is assumed to be the most current and accurate representation and will be assessed for accuracy. However, it is preferable to work from any original line format version that might be available.

Fort Hood Installation Boundary Data
Initial Data Set, April1999

Map Layout #1



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Training Areas Data (See Map Layouts #2 and #3)

Five applicable data sets in FTP transfer:

Trnareas.shp	290 polyline features
Trnafull.shp	445 polyline features
Trnafullp.shp	140 polygon features
Livefire.shp	39 polyline features
Pd94.shp	17 polyline features

A sixth file was delivered in a later transfer:

Trnafull.shp	389 polyline features

Trnareas contains polyline features for the main training areas (e.g., areas 1,2,3) and does not distinguish the subareas (e.g., areas 3a and 3b). Attributes in the data table indicate the data came from a CAD source:

layer (AV_TRN_AREAS or blank)
elevation (all set to 0.00000)
thickness (all set to 0.00000)
color (if layer = AV_TRN_AREAS then 1 else 0)
length (ranges from 6.808 to 4851.112; measure unknown)

FIfty-eight of the 290 polylines have no layer or color. This may be an indication that they were an update to the original file, perhaps after it was transferred to ArcView.

The originally delivered *Trnafull* contains polyline features for the main training areas and subareas. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 0 to 297)
tnode_ (ranges from 0 to 297)
lpoly_ (all set to 0)
rpoly_ (all set to 0)
length (ranges from 5.186 to 30288.624; measure unknown)
trnafull_ (ranges from 0 to 435)
trnafull_i (ranges from 0 to 410)

Thirty-six of the 445 polylines have zero values for the typical ARC/INFO topologic and id fields. This may be an indication that they were an update to the

original file, perhaps after it was transferred to ArcView. A visual comparison of these unattributed polylines with those from *trnareas* shows that they are boundaries for the same areas, but were independently updated (digitized).

Trnafullp contains polygon features for the main training areas and subareas. Attributes in the data table indicate the data came from a Grass source:

Seq_num (1 to 146, no duplicates, a few breaks in the sequence)
Grass_area (0 to 136)
Cat (typical Grass attribute designation, string field at least 16 characters)
Training_a (string field, similar to Cat)

For 132 features, the notation for Cat field begins with "TA" followed by a numeric/alpha designation (e.g., "TA 14B" or "TA 20"). Usually the notation for training_a matches that of Cat except that the "TA" prefix is dropped. One exception is for Cat="TA 14A", where training_a="PHANTOM RUN". The remaining eight features have distinct notations (e.g., "BLORA", "Contonment (old 9)", "Robert Gray AAF").

Eleven of the 140 polygons have zero values for the Grass area field, and range in seq_num from 136 to 146. This may be an indication that they were an update to the original file, perhaps after it was transferred to ArcView. A visual comparison of these unattributed polygons with those from *trnafull* shows that they are often defining the same areas, but were independently updated (digitized).

There are two instances where one polygon completely overlays a second polygon. The first example is in the Dudded Area; the polygon defined as "Historic Dudded Area" is completely contained within the polygon defined as "Permanently Dudded Area". The second example is with the Army Air Field; the polygon defined as "TA 24 C" is completely contained within the polygon defined as "Robert Gray AAF".

Livefire contains polyline features defining a single boundary central to the installation. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 2 to 37)
tnode_ (ranges from 1 to 38)
lpoly_ (all set to 0)
rpoly_ (all set to 0)

length (ranges from 154.203 to 4344.383; measure unknown)
livefire_ (ranges from 1 to 38)
livefire_i (ranges from 1 to 38, values sometimes match values of livefire_)

A spatial comparison between *livefire* and *trnafull* found matches for 30 of the 39 features. Attributes matched only on the length field. The nine unmatched lines are not related to the presumed changes in *trnafull* after it was moved to ArcView.

Pd94 contains polyline features defining another, smaller single boundary central to the installation. Attributes in the data table provide no indication of the data's origins:

id (all set to 0)
length (ranges from 514.756 to 3192.951)

A spatial comparison between *pd94* and *trnafull* found zero matches. A spatial comparison between *pd94* and *trnareas* found matches for 11 of the 17 features. Attributes matched only on the length field. All the lines in *pd94* are similar to some of the presumed changes in *trnareas* after it was moved to ArcView.

The replacement *Trnafull* contains polyline features defining the major training areas and subareas. No description accompanied this file. The attribute data is very similar to the originally delivered *trnafull*, having the identical attribute names and value ranges:

fnode_ (ranges from 0 to 297)
tnode_ (ranges from 0 to 297)
lpoly_ (all set to 0)
rpoly_ (all set to 0)
length (ranges from 0.0 to 30288.624; measure unknown)
trnafull_ (ranges from 0 to 435)
trnafull_i (ranges from 0 to 410)

Also like the original trnafull, the replacement data set has features (22) with zero values for the typical ARC/INFO topologic and id fields. This may be an indication that they were an update to the original file, perhaps after it was transferred to ArcView.

A spatial comparison between the original and replacement *trnafull* data sets found only 22 matches, all for features along the eastern installation boundary at the reservoir. A visual inspection confirmed that the same areas are generally

being defined in both data sets, but the geometries of the area boundaries are very different. A table join of the id attributes (a join of attributes trnafull_ from the two sets) found 367 matches. Exactly 261 of these matching features also have identical length values. Further investigation showed:

1. Features that match on id generally represent the same area boundary, but with different geometries.
2. Features that do not match on id are generally the result of two or more features being combined into one in the replacement data set, but often in error because lines are no longer connecting correctly at nodes.

All these comparisons indicate that the two data sets were likely derived from the same parent or one from the other, but that the line geometries were significantly changed. The missed nodes is an issue if we want to transform the line features into polygons. The length field is suspect in light of the significant alterations in geometry, and should be recalculated.

The boundaries of the outermost training areas seem to coincide with the installation boundary defined in the replacement *Universe* data set. Spatial comparison cannot be performed because the feature types do not match. However, a close visual inspection shows that the geometry of the training area boundaries and the installation boundary are very similar but not identical. Thus it is likely that this data set was developed with the same methodology and from the same source as the replacement *universe*.

In addition to the 6 files examined, the CD referenced over 60 other data sets that represent subsets of the training areas. These may or may not have been generated as extracts of one of the overall training area data sets.

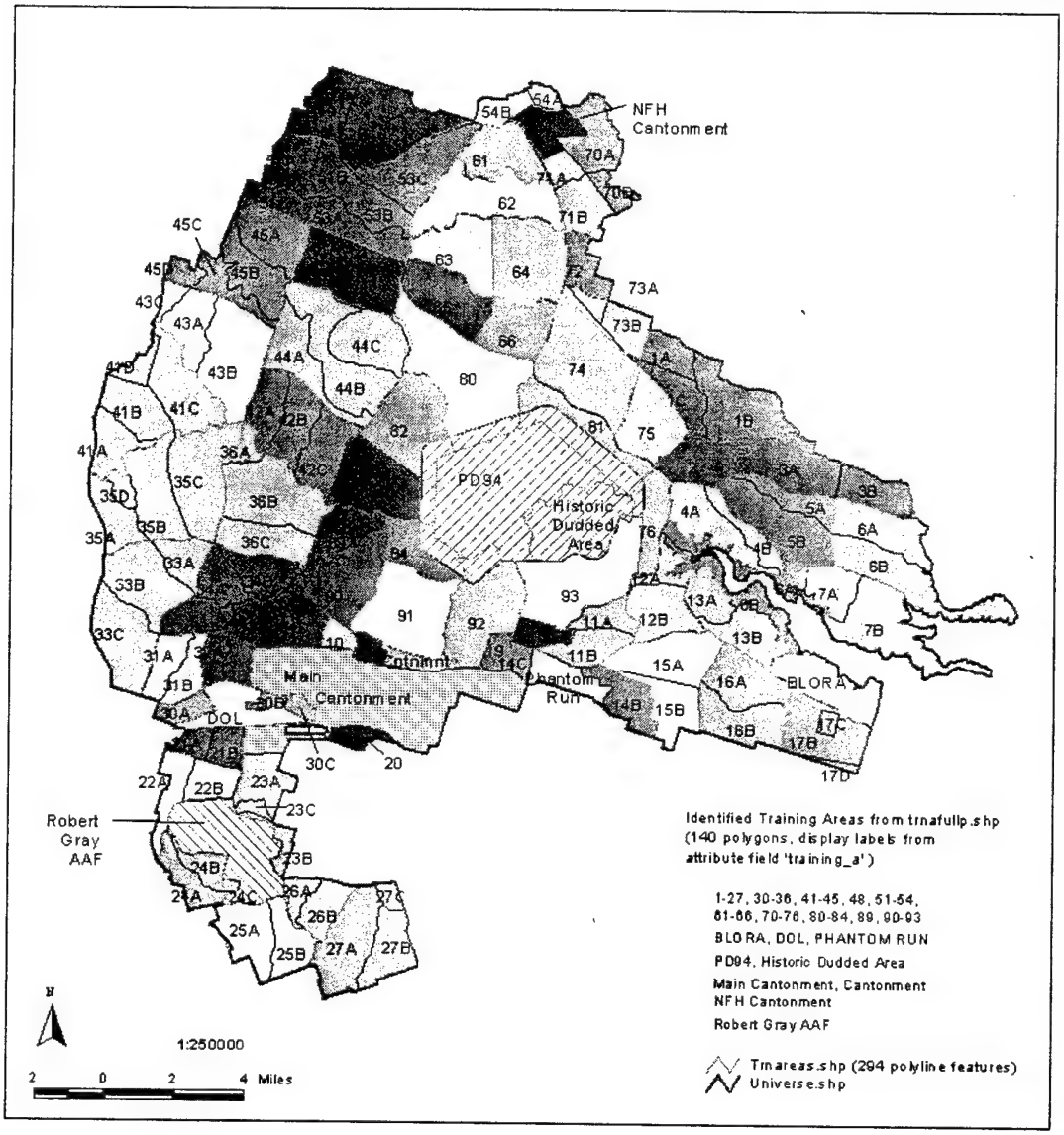
All the data sets have slight topologic differences between them, indicating that each may have a slightly different spatial accuracy (see Layout #3). Generally the data sets are very similar, however. *Trnafullp* is the only data set populated with descriptive labels. This could be used in conjunction with the replacement *trnafull*, which is assumed to be the most current.

A comparison was made between the boundaries represented in the training areas data sets and the roads and streams data sets. These data sets were compared to determine if spatial accuracy of training area boundaries could be measured by comparison to these easily identified features. It appears that training area boundaries are often defined by roads, but also may be defined by streams, tank trails, or other unidentified linear features. However, no specific guidance on how the training area boundaries are delimited is available.

Fort Hood Training Areas

Initial Data Set, April 1999

Map Layout #2



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

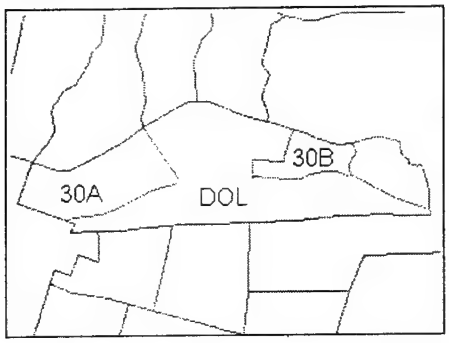
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Training Areas - Details

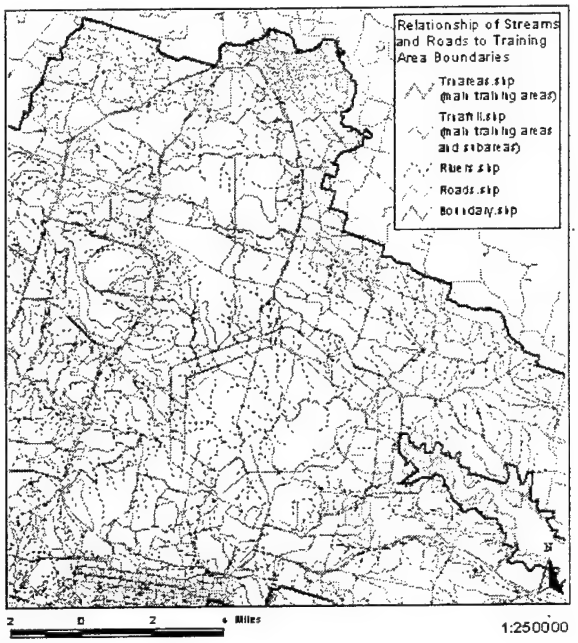
Initial Data Set, April 1999 Map Layout #3

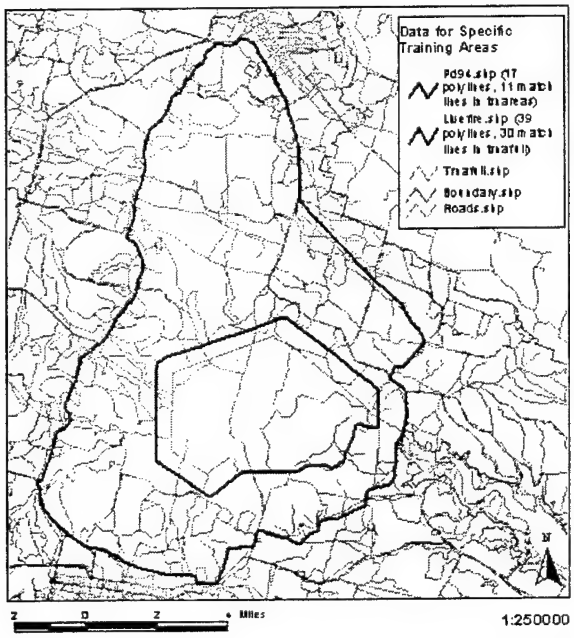
Trnafull.shp 445 polylines



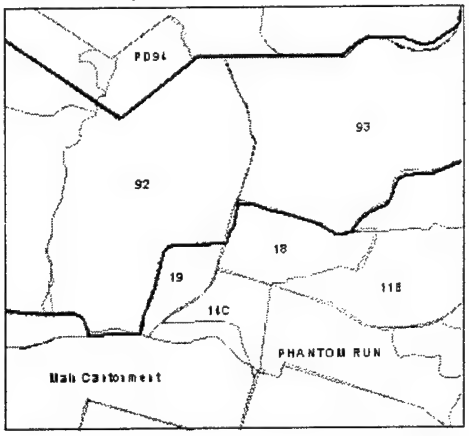
1:100000

Data appears to have originated in Arc/Info (data has typical topology fields). Data appears to have been modified after transfer to ArcView. 36 lines (examples in blue) have no values in topology fields and do not connect at nodes in original data.





Detail comparison of topology for all training area boundary files



1:100000

Pd94.shp
Livefire.shp
Trnafull.shp
Trnareas.shp
Trnafullp.shp

Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Surface Hydrology Data (See Map Layouts #4 through #7)

Eight applicable data sets in FTP transfer:

alldam97.shp	377 polygon features
lakes.shp	187 polyline features
ponds.shp	155 point features
riverall.shp	2698 polyline features
riverlg.shp	640 polyline features
riverlgall.shp	284 polyline features
rivers.shp	540 polyline features
riversm.shp	1268 polyline features

A ninth file was delivered in a later transfer:

Alldam99bnd.shp	218 polygon features

Alldam97 contains polygon features for the reservoirs adjacent to and southeast of the installation. Some 179 features are within the boundary of the installation. Attributes in the data table indicate the data probably came from an ARC/INFO source, but has been modified:

area (ranges from 72.323 to 197903.793, measure unknown)
perimeter (ranges from 40.174 to 4651.002, measure unknown)
alldams96_ (ranges from 0 to 81, 339 with zero value)
alldams96_ (all set to 0)
name (one feature each set to: "TA72-1","TA73-1","TA73-2"; the remainder "")
remarks (same features with name have various comment, e.g., "Dry During Summer"; the remainder "")
hectares (ranges from 0.007 to 19.783)

The additional *alldam99bnd* is very similar to *Alldam97*. The following description accompanied this file:

> New ponds/dams layer - with correct attributes - some GPS collected, most heads-up digitized from 1997 imagery - layer clipped to Universe layer.

The attribute data is comparable to *alldam97*, having identical attribute names and similar value ranges:

area (ranges from 134.409 to 197903.793, measure unknown)
perimeter (ranges from 48.590 to 4651.002, measure unknown)
alldams96_ (ranges from 0 to 75, 188 with zero value)
alldams96_ (ranges from 0 to 54, 192 with zero value, not same as previous attribute)
name (36 named (i.e., "Nolan Lake"), 148 labeled by training area (i.e., "31A", "31B",... "31E"), 31 blank or question marked)
remarks (open ended text field, usually describes size, permanence)
hectares (ranges from 0.013 to 19.777)

A spatial comparison between *alldam97*(179) and *alldam99bnd* found 141 matches. Some of the 39 mismatches can be explained by small revisions to a features geometry or location, while others may have been erroneously included in the original file. A*lldam99bnd* also contains a number of features that were not represented in *alldam97.*

Ponds contains point features presumably indicating the location of ponds within the installation. It is not clear what the point actually references; for example, the center of the pond or the discharge location. Attributes in the data table indicate the data came from an ARC/INFO source, but that the data was converted from a polygon to a point form. Most of the fields are 0 value:

area, perimeter (all set to 0)
ponds_, ponds_id (ranges from 1 to 155, values for both attributes same for each feature)
data (all set to blank)

The description provided with *alldam99bnd* implied that the data set included ponds and dams. To assess overlap, a spatial comparison was made between *alldam99bnd* and *ponds.* Some 15 of the pond features were located within the bounds of an *alldam99bnd* polygon. An additional 66 were within 200 (units unspecified) of an *alldam99bnd* polygon. It is unclear whether some or any of the features of ponds are represented in *alldam99bnd.*

Lakes contains polyline features defining a set of water bodies throughout the installation. The attributes in the data table indicate that data probably came from DLG data source. Since the 1:24000 DLGs have not been produced for this area, it can be concluded that this lake's data was produced from 1:100,000:

ID (ranges from 4 to 159)
Major1 (all set to 50)
Minor1 (all set to 200)
Major2 (all set to 0)
Minor2 (all set to 0)
Major3 (all set to 0)
Minor3 (all set to 0)

The *lakes* data set was compared to the hydrography data from downloaded USGS DLG data for the Fort Hood area. The DLG hydrography data consists of both lakes and stream features. This DLG data was imported into ARC/INFO, and transformed from NAD27 to NAD83 to make it compatible with Fort Hood spatial data. While the DLG lake features visually seemed to match with the Fort Hood *lakes* data set, a spatial join found no exact matches. A detailed manual inspection indicated that the spatial join failed because the features are shifted by approximately 1 meter, probably caused by different projection conversion processes.

Another version of *lakes.shp* was delivered on the compact disc. This data set is separately reviewed here because it is different from but appears closely related to the ftp version:

Lakes.shp 23 polygon features

The polygon *lakes* contains features only for the reservoirs adjacent to the southeast section of the installation. Attributes in the data table indicate the data came from an ARC/INFO source:

area (ranges from 0.0 to 47270000.0, measure unknown)
perimeter (ranges from 0.0 to 210700.0, measure unknown)
lakes1_ (ranges from 0 to 23, no zero value feature)
lakes1_id (all set to 0)

Two of the 23 polygons account for 99% of the total area represented by the data set. A detailed visual inspection of the map shows that the original source was probably a grid or raster representation that was translated to polygon form

before transfer to ArcView. The separation into 23 polygons is likely an artifact of the various translation processes.

Riverall contains polyline features defining an extensive stream network related to the installation. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 8 to 2637)
tnode_ (ranges from 1 to 2637)
lpoly_, rpoly_ (all set to 0)
length (ranges from 1.288 to 31738.573; measure unknown)
riversall_ (ranges from 1 to 2698)
riversall_ (ranges from 1 to 2680)

The data set represents major rivers both inside and outside the installation boundary, likely from their source to their mouth. Smaller rivers/tributaries are included if they have any connection to the installation. Each stream is represented as a single polyline, assumed to be the centerline. No attributes indicate a relationship between polylines to form a continuous stream.

The data set also contains polylines representing surface water extents for water bodies connected to the stream network. At these locations the stream is represented as a linear feature bisecting the water body. These water bodies often coincide with, but are not identical to, the areas defined in the polyline *lakes* data set.

Riverlg contains polyline features defining a subset of the stream network within the installation boundary. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 0 to 644)
tnode_ (ranges from 0 to 643)
lpoly_, rpoly_ (all set to 0)
length (ranges from 7.906 to 6472.401; measure unknown)
riverlg_ (ranges from 0 to 640)
riverslg_id (ranges from 0 to 638)

The data set apparently represents major rivers inside installation boundary. There are some discrepancies in the data where segments are missing, mostly segments that defined the river's path across the installation boundary.

A spatial comparison between *riverlg* and *riverall* found 569 matches. Some of the 71 mismatches can be explained: features crossing the installation boundary were clipped and therefore do not match the presumed original from *riverall*. However, some of the unmatched features do represent changes in the location/path of the feature.

Riverlgall contains similar polyline features but defines a second subset of the stream network both inside and outside the installation boundary. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 3 to 280)
tnode_ (ranges from 1 to 280)
lpoly_, rpoly_ (all set to 0)
length (ranges from 22.406 to 31738.573; measure unknown)
riverslg_ (ranges from 1 to 284)
riverslg_i (ranges from 0 to 638)

The data set apparently represents major rivers both inside and outside the installation. Since there are only three main rivers and a small selection of tributaries in this data set, the definition of what qualifies as a "large" river is much stricter than in *riverlg*.

A spatial comparison between *riverlgall* and *riverall* found 279 matches. The five unmatched features represent minor changes in the location/path of the feature probably made to correct a problem identified by the removal of connecting features.

Rivers contains polyline features defining a stream network in and around the installation, and also includes some water bodies. The attributes in the data table indicate the data probably came from the 1:100000 USGS DLG data:

ID (ranges from 4 to 159)
Major1 (all set to 50)
Minor1 (all set to 200)
Major2 (all set to 0)
Minor2 (all set to 0)
Major3 (all set to 0)
Minor3 (all set to 0)

The *rivers* data set was compared to the hydrography data from downloaded USGS DLG data for the Fort Hood area. While the DLG features visually seemed to match with the Fort Hood *rivers* data set, a spatial join found no exact

matches. A detailed manual inspection indicated that the spatial join failed because the features are shifted by approximately 1 meter, probably caused by different projection conversion processes.

A spatial comparison between *rivers* and the polyline *lakes* data set found a match for every feature in *lakes*.

As with lakes, a different version of *rivers.shp* was delivered on the compact disc:

Rivers.shp1 712 polyline features

This version of *rivers* contains polyline features defining an extensive stream network within the installation. Attributes in the data table indicate the data came from a CAD source:

entity (all set to "polyline")
layer (43 features set to "AV_RIVERS_BND", the remainder "LV_RIVERS_BND")
elevation (all set to 0.00000)
thickness (all set to 0.00000)
color (if layer = AV_RIVERS_BND then 1 else 2)

The extents of this river network seem to match with *boundary.shp* rather than *universe.shp*, evidenced by some missing segments where the river passes outside the installation at the reservoir.

The features defined as "AV_RIVERS_BND" represent small areas at the south end of the installation, on/around some linear features in the data set. These areas often correspond to features in *alldam97*, though not all features in *alldam97* are represented in *rivers*.

A spatial comparison between *rivers* and *riverall* found no matches. However, a detailed manual inspection found that these features differ locationally by 1 or 2 inches. This slight shift may have been caused by a different approach to projection or storage type conversion.

Riversm contains polyline features defining a subset of the stream network within the installation boundary. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 0 to 1663)
tnode_ (ranges from 0 to 1662)

lpoly_, rpoly_ (all set to 0)
length (ranges from 0.0 to 4005.820; measure unknown)
riversm_ (ranges from 0 to 1264)
riversm_i (ranges from 0 to 1256)

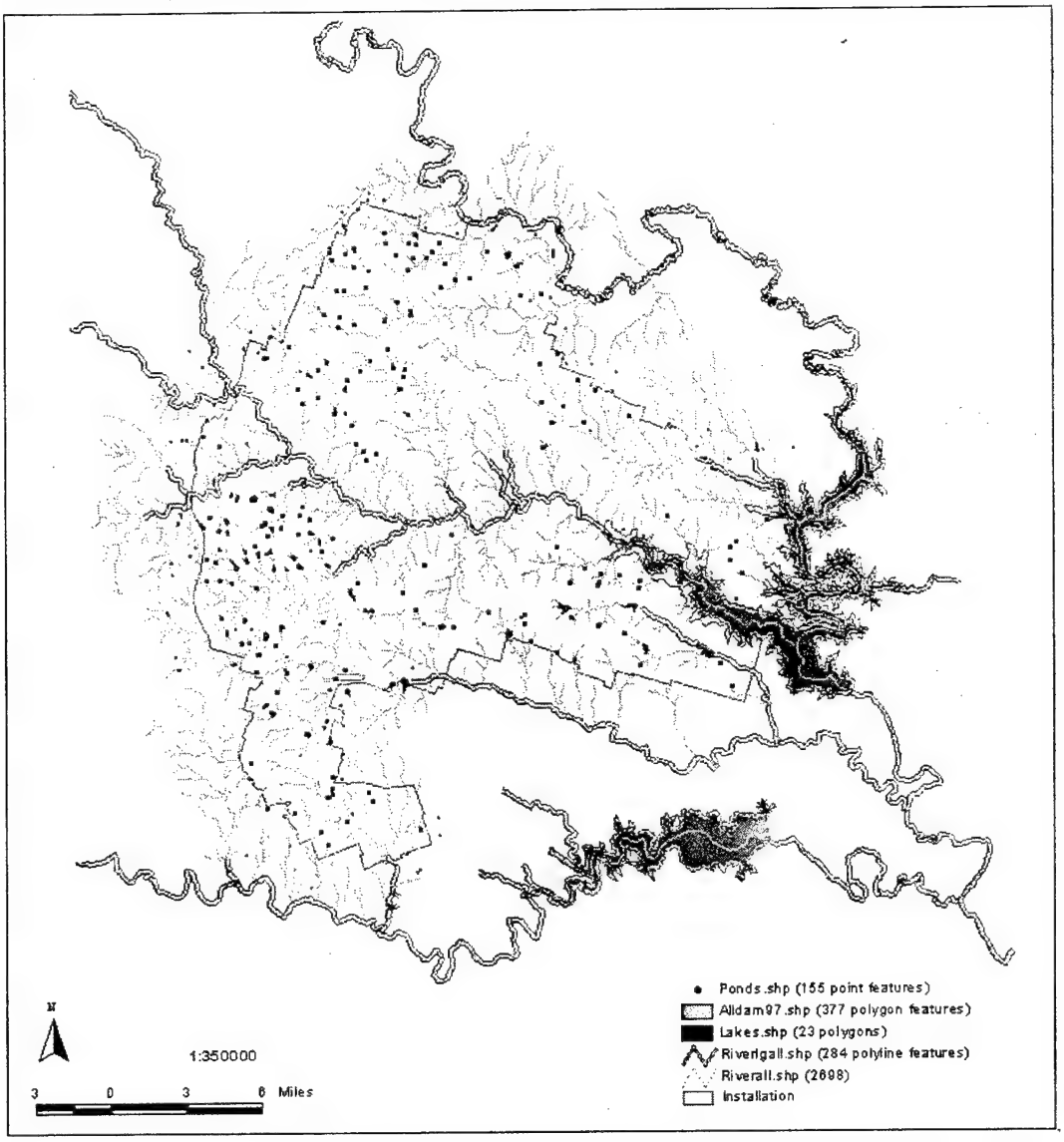
The data set apparently represents tributaries and feeder streams inside the installation.

A spatial comparison between *riversm* and *riverall* found 1095 matches. The 173 unmatched features may represent minor changes in the location/path of the feature probably made to correct a problem identified by the removal of connecting features.

It is not clear how the polylines were identified as "large" or "small" for inclusion/exclusion from the various subsets of river data. There are 1446 features from *riverall* that are not represented by either of the subset data sets. Though many are outside the installation boundary, many are within the boundary, widely distributed across the installation, and connect to other features that were classified. Also, there are 177 features that are included in both the "large" and "small" data sets.

Fort Hood Surface Hydrology - Lakes, Ponds, Dams, and River Extents

Initial Data Set, April 1999 Map Layout #4



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

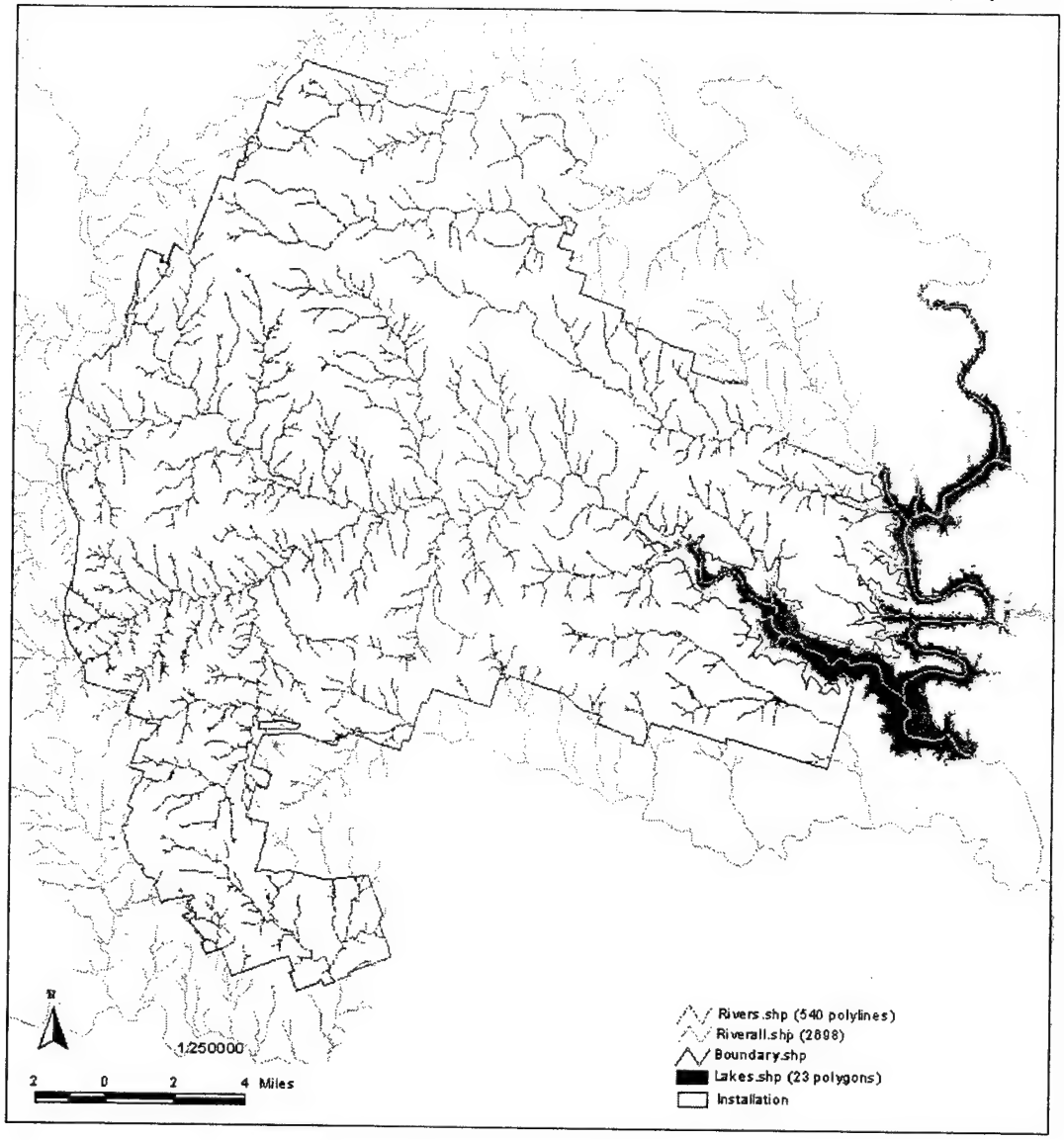
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Surface Hydrology - Comparison of Complete Networks

Initial Data Set, April 1999

Map Layout #5



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

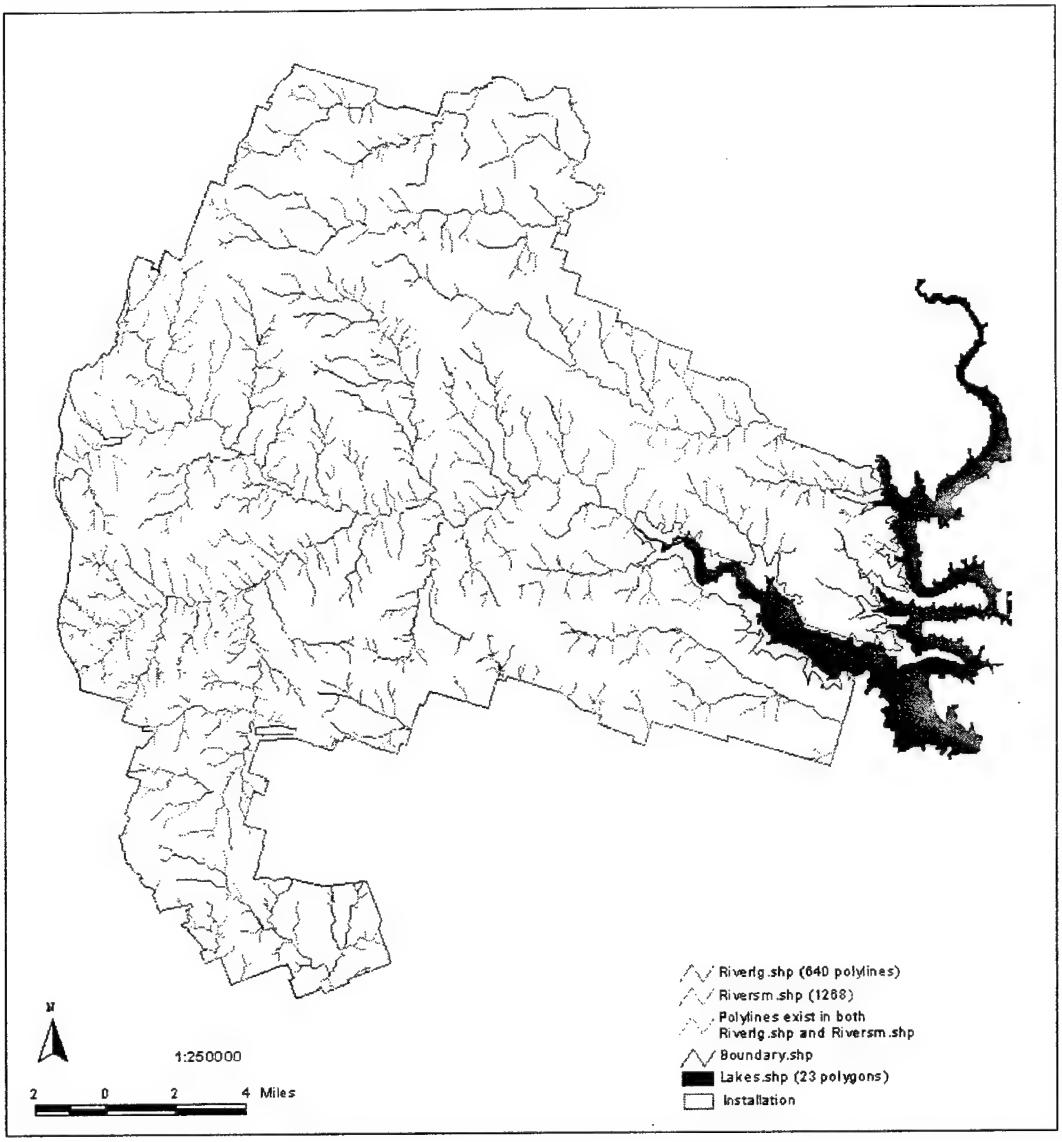
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Surface Hydrology - Comparison of Size Classifications

Initial Data Set, April 1999

Map Layout #6



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

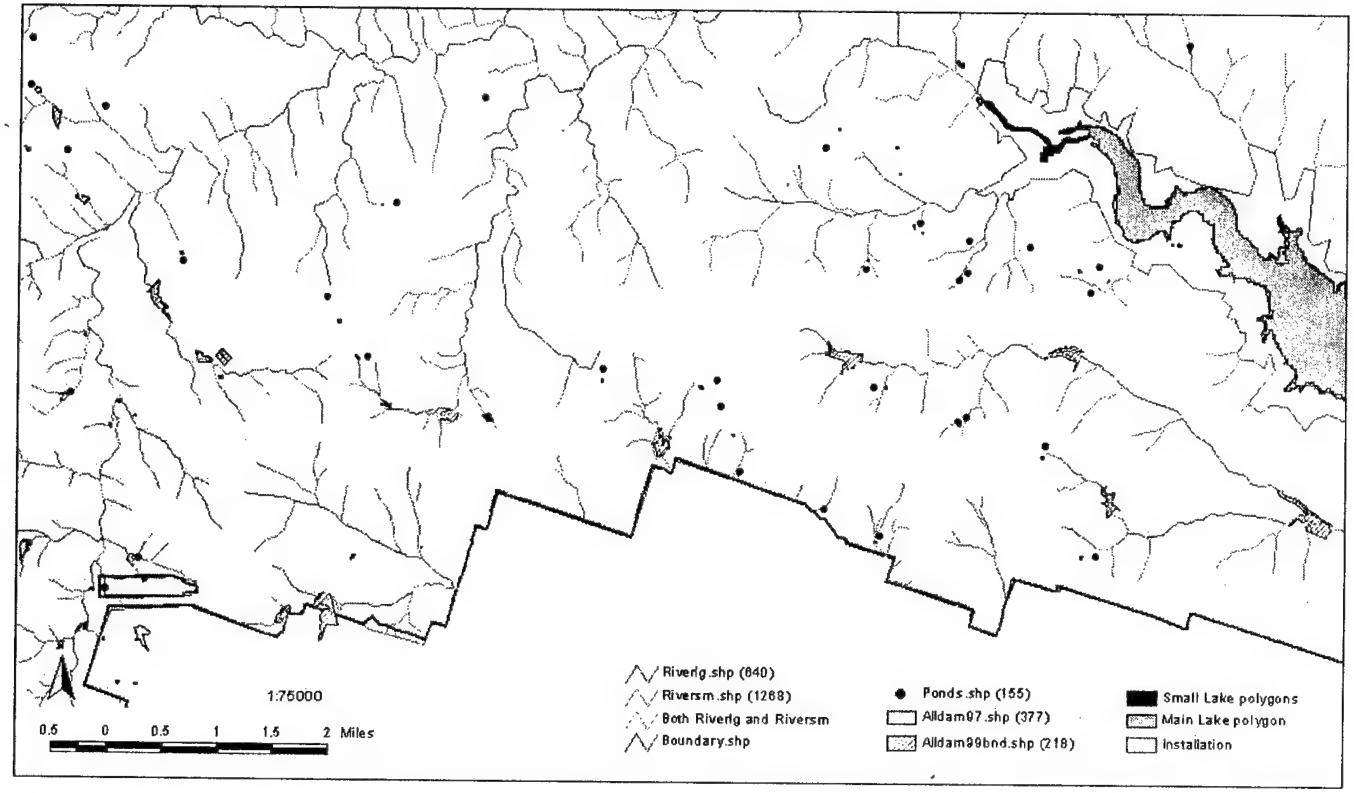
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Surface Hydrology Detail

Initial Data Set, April 1999

Map Layout #7



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Stream and Pipeline Crossing Data (See Map Layouts #8 through #11)

Three applicable stream crossing data sets in FTP transfer:

Lowwaterxing.shp	13 point features
Strmxing.shp	20 point features
Strmxpro.shp	74 point features

Lowwaterxing contains unattributed point features:

id (all set to 0)

Four features from the data set occur along the perimeter of the live fire area (*livefire*), the remaining nine to the west and southwest of the live fire area. The ITAM spatial database documentation describes this file as "supplied by Fort Hood Game Warden Office, Oct. 1999." Nothing indicated how the data was collected and recorded or how current it is.

Strmxing contains point features with attributes offering little distinctive information about the data:

area, perimeter (typical polygon attributes, all set to 0)
strmxing_, strmxing_I (ranges from 1 to 20, identical values in both fields)
data (all set to "Hardened Stream Crossing Site")

According to the documentation, this data set contains "medium water crossings." Five features from the data set occur along the perimeter of the live fire area (*livefire*), one within this area, and two to the east. The remaining 12 are distributed throughout the western section of the installation.

Strmxpro contains identical attributes to *strmxing*, but with minor changes in values:

area, perimeter (typical polygon attributes, all set to 0)
strmx_prop, strmx_prop (ranges from 1 to 74, identical values in both fields)
data (all set to "#2")

According to the documentation, this data set contains "proposed stream crossing sites, LRAM, 1997". The sites are distributed throughout the western section of the installation, north of Copperas Cove Road.

A spatial join between *lowwaterxing* and *strmxing* found no duplicate features. However, five features matched within 50 to 75 meters, two within 25 meters. A spatial join between *strmxing* and *strmxpro* found no proximate matches (distances were mostly greater than 1000 meters).

There were no data files related to pipelines or pipeline crossings in the FTP transfer. However, there were two files from the CD:

Pipeline.shp	68 polyline features
Pipelbnd.shp	27 polyline features

A third file was delivered in a later transfer:

Pipexing.shp	21 polyline features

Pipeline contains polyline features presumably defining the pipelines in and around the installation. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 0 to 85)
tnode_ (ranges from 0 to 84)
lpoly_, rpoly_ (all set to 0)
length (ranges from 0.0 to 15444.567; measure unknown)
pipeline_ (ranges from 0 to 63)
pipeline_i (ranges from 0 to 63)

Five features from the data set have zero values for all the attributes. This may indicate that they were an update to the original file, perhaps after it was transferred to ArcView.

Pipelbnd contains polyline features defining only those pipelines within the installation boundary. Attributes are identical to *pipeline*:

fnode_ (ranges from 1 to 33)
tnode_ (ranges from 2 to 32)
lpoly_, rpoly_ (all set to 0)
length (ranges from 9.494 to 12311.690; measure unknown)
pipeline_ (ranges from 1 to 27)
pipeline_i (ranges from 1 to 27)

A spatial comparison between *pipelbnd* and *pipeline* found 19 matches. The five 0 length features from *pipeline* matched features that had values for length in

pipelbnd. Otherwise attributes matched on the length field but not on any other attribute.

The eight unmatched features mostly represent differences caused by clipping lines that extended across the boundary, although at least two are minor differences in length and position caused by topology changes.

The USGS DLG data for the Fort Hood area contained a data set for pipeline features. This data was imported into ARC/INFO and converted from NAD27 to NAD83. A spatial join between the DLG and Fort Hood data found no exact matches. However, a detailed manual inspection indicated that the Fort Hood data likely originated from the DLG data, but has undergone editing. The two data sets contain many features within the installation boundary which match in length and geometry. The spatial join failed because these features are shifted by approximately 0.5 meters, probably caused by different datum conversion processes. Most data set differences can be explained as deletions to remove the errors, a scattering of small discontinuous features, in the original DLG data. The additional features in the Fort Hood pipeline data set are the features with zero values for all attributes.

Unlike the stream crossings, *pipexing* contains polyline features. The lines presumably represent the centerline and length of approved pipeline crossings. Attributes in the data table indicate the data came from an ARC/INFO source:

fnode_ (ranges from 1 to 41)
tnode_ (ranges from 3 to 42)
lpoly_, rpoly_ (all set to 0)
length (ranges from 260.308 to 536.026; measure unknown)
pipexing_ (ranges from 1 to 21)
pipexing_i (ranges from 1 to 21, no matches w/ pipexing_)

All the features from the data set occur along the northern pipeline. Five features appear within the livefire area, 10 to the east, and 6 to the west. Five of the features do not correspond to features from *roads.shp*, but the remaining 16 are good or proximate matches.

The following files transferred on the compact disc seem to represent additional information about crossings:

Crossing.shp	18 point features
Fy98xing.shp	9 point features
Fy99xing.shp	9 point features

Lramstrx.shp	(duplicate of strmxpro)
Prostrmx.shp	692 point features
Streamx33.shp	6 point features
Strmx33.shp	(duplicate of streamx33)
Tempxing.shp	17 point features

Crossing contains point features including some descriptive attributes about the features:

area, perimeter (typical polygon attributes, all set to 0)
crossings_, crossings_ (ranges from 1 to 18, identical values in both fields)
stream_vel (ranges from 0.0 to 0.920; 13 features set to 0.0)
width_of_s (ranges from 0.0 to 10.1; 8 features set to 0.0)
bottom_mat ("Gravel and Sand", "NA", "Rock and Mud")
depth (0.0, 0.2, 0.3, 0.4, 0.50; 9 features set to 0.0)
type ("Ford", "NA")
stream ("Cottonwood Creek","Cowhouse Creek","Table Rock Creek")

All features with width_of_s set to 0 have no information for attributes stream_vel, bottom_mat, depth, type (value set to 0 or NA). All features have a stream (name) attribute and occur west of the live fire area along a common section of the stream network.

A spatial join between *crossing* and *strmxing* found three proximate matches (distances less than 25 meters); the remaining 15 were all in excess of 290 meters. A spatial join between *crossing* and *strmxpro* found only one possible match (distance less than 100 meters); the remaining 17 ranged from 100 to 1500 meters away from the closest feature in *strmxpro*.

FY98xing contains attributed point features:

id (all set to 0)
name ("Georgetown Crossing", "Ripstein #3", "Ripstein #4", "Table Rock #1", "Table Rock #2", blank)
remarks ("6/6/98","EOY98",blank)

Two features of *fy98xing* occur on the same section of the stream network as *crossing*; the remaining seven occur on the section just south. A spatial join between these data sets showed that they may match on one feature, but the distance (349 meters) may indicate they reference different crossings.

FY99xing contains unattributed point features:

id (all set to 0)

All features of *fy99xing* occur on the same section of the stream network as *crossing*. A spatial join between *fy99xing* and *crossing* found 3 probable matches (distances less than 200 meters). A spatial join between *fy99xing* and *fy98xing* found only one possible match (350 meters distance), but a visual inspection shows the two points likely reference different streams.

Prostrmx97, like *strmxing,* contains point features with attributes, but offering little distinctive information about the data:

area, perimeter (typical polygon attributes, all set to 0)
prostrmx_, prostrmx_ (ranges from 1 to 692, identical values in both fields)
data (all set to "#0")

All features of *prostrmx97* occur west of the live fire area and north of Route 190.

A spatial join between *prostrmx97* and *strmxpro* found the 56 of *strmxpro's* 74 features are likely matches to features in *prostrmx97* (distance less than 200 meters). A visual inspection of the 12 features with nearest distances of more than 200 meters shows that many of these are likely matches to prostrmx97 features because of locational discrepancies (not actually positioned on a stream).

A spatial join between *prostrmx97* and *strmxing* found eight probable (less than 100 meters) and four possible (between 100 and 200 meters) matches. A spatial join between *prostrmx97* and *crossing* found 13 probable and 3 possible matches.

Streamx33 contains typical attribute data, and appears to represent stream crossings in training area 33:

area, perimeter (typical polygon attributes, all set to 0)
streams33_, streamx33_ (ranges from 1 to 6, identical values in both fields)

This data set doesn't appear to be a direct subset of any of the other data sets. Four features are probable matches for features from *fy98* (less than 30 meters distance). One of the remaining two features is probably incorrectly located (it is about 400 meters from the nearest stream), the other may match a feature in *prostrmx*.

Tempxing, like *Lowwaterxing,* contains unattributed point features:

id (all set to 0)

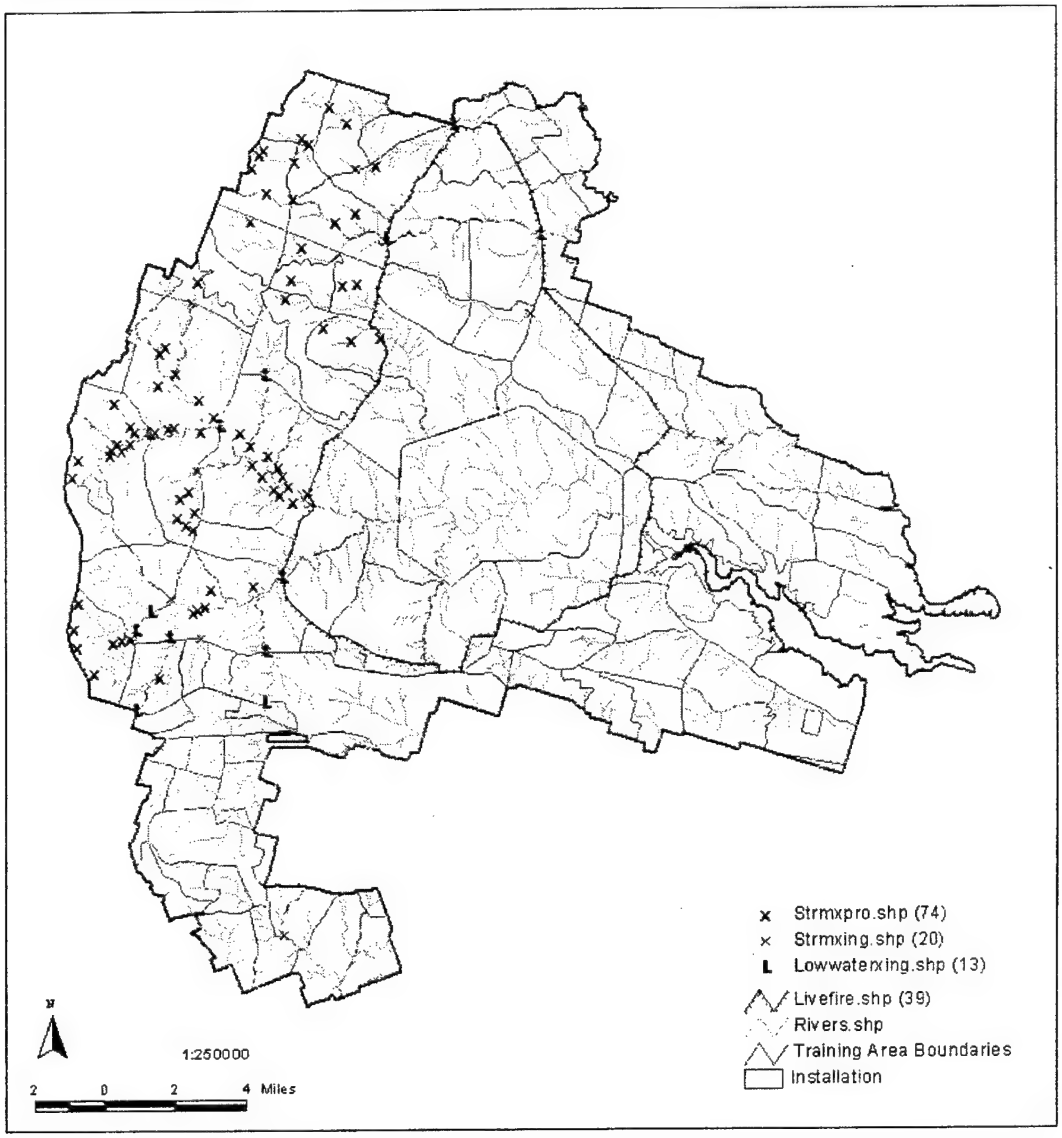
All features from the data set occur along the western boundary of the installation. A spatial join between *tempxing* and *strmxing* found no proximate matches (distances all in excess of 2000 meters). A spatial join between *tempxing* and *strmxpro* found two proximate matches (distances less than 200 meters).

Prostrmx97 appears to be the most complete in terms of number of stream *crossing* identified. *Crossing* contains the most detailed set of attributes but for a limited number of features. None of the data sets seem to represent pipeline crossings, simply because none of their features are proximate to the pipelines as represented by *pipelbnd*.

Fort Hood Stream Crossing Data

Initial Data Set, April1999

Map Layout #8



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

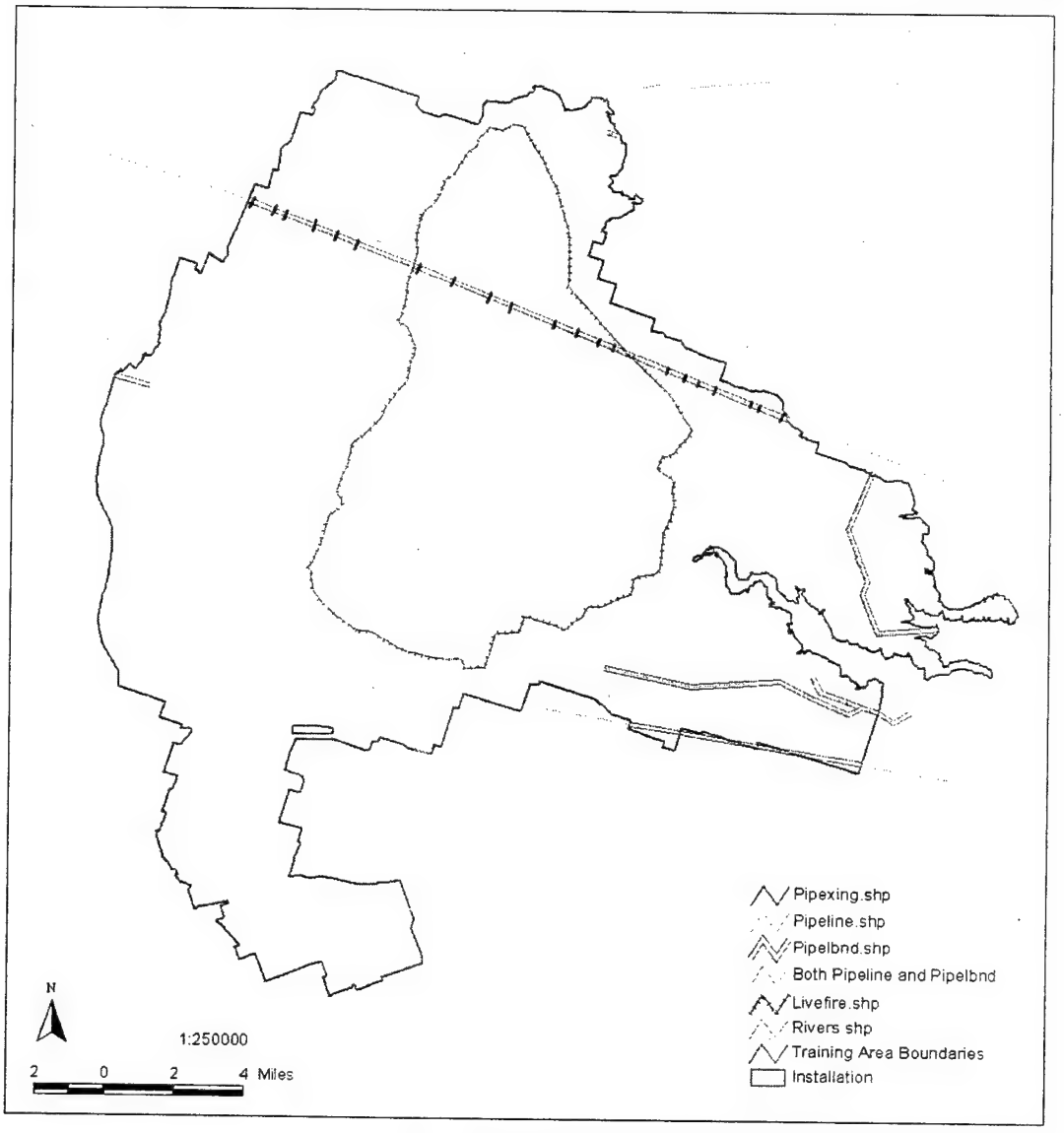
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Pipeline Crossing Data

Initial Data Set, April1999

Map Layout #9



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

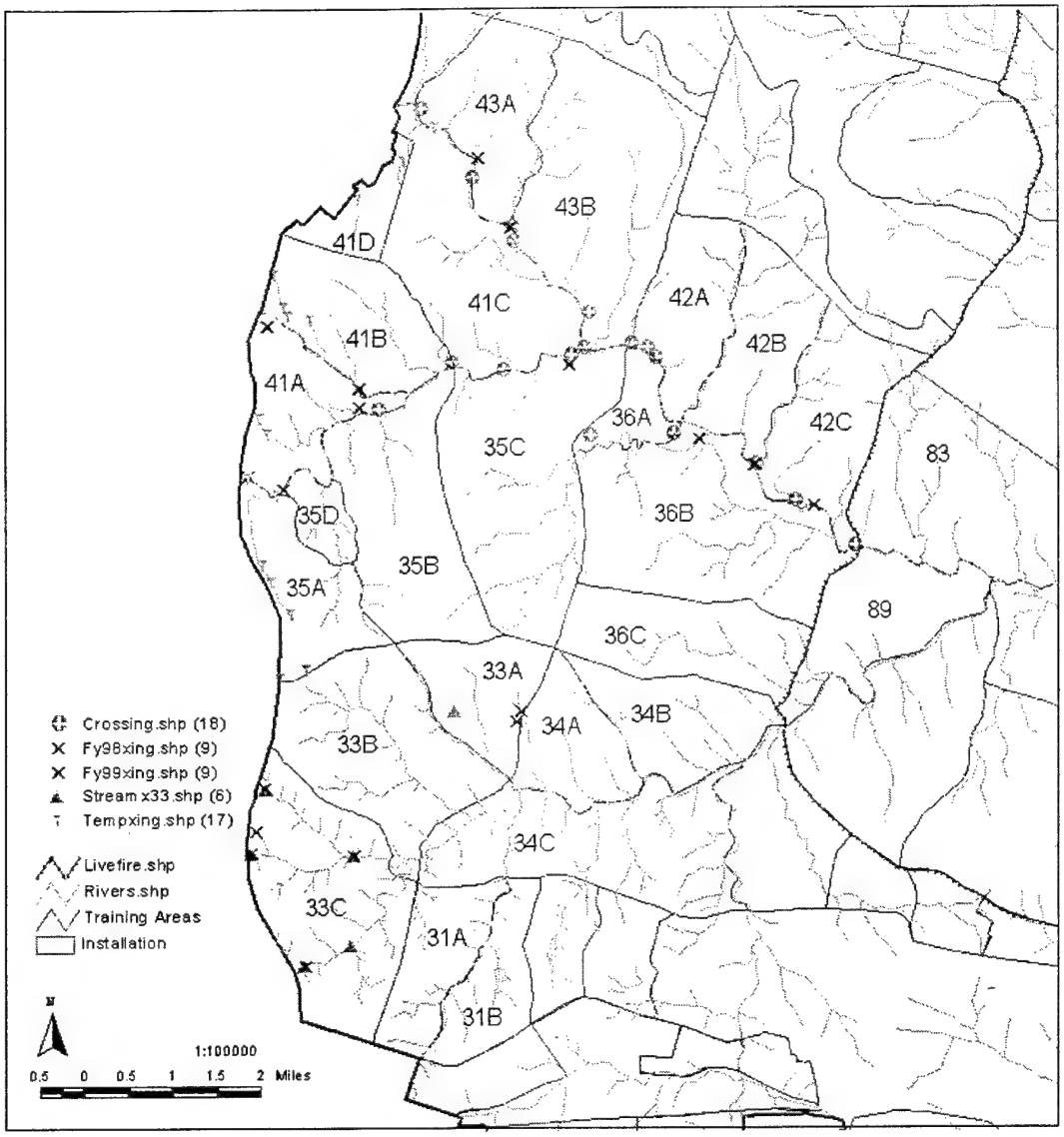
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Stream Crossing (Additional Data)

Initial Data Set, April 1999

Map Layout #10



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

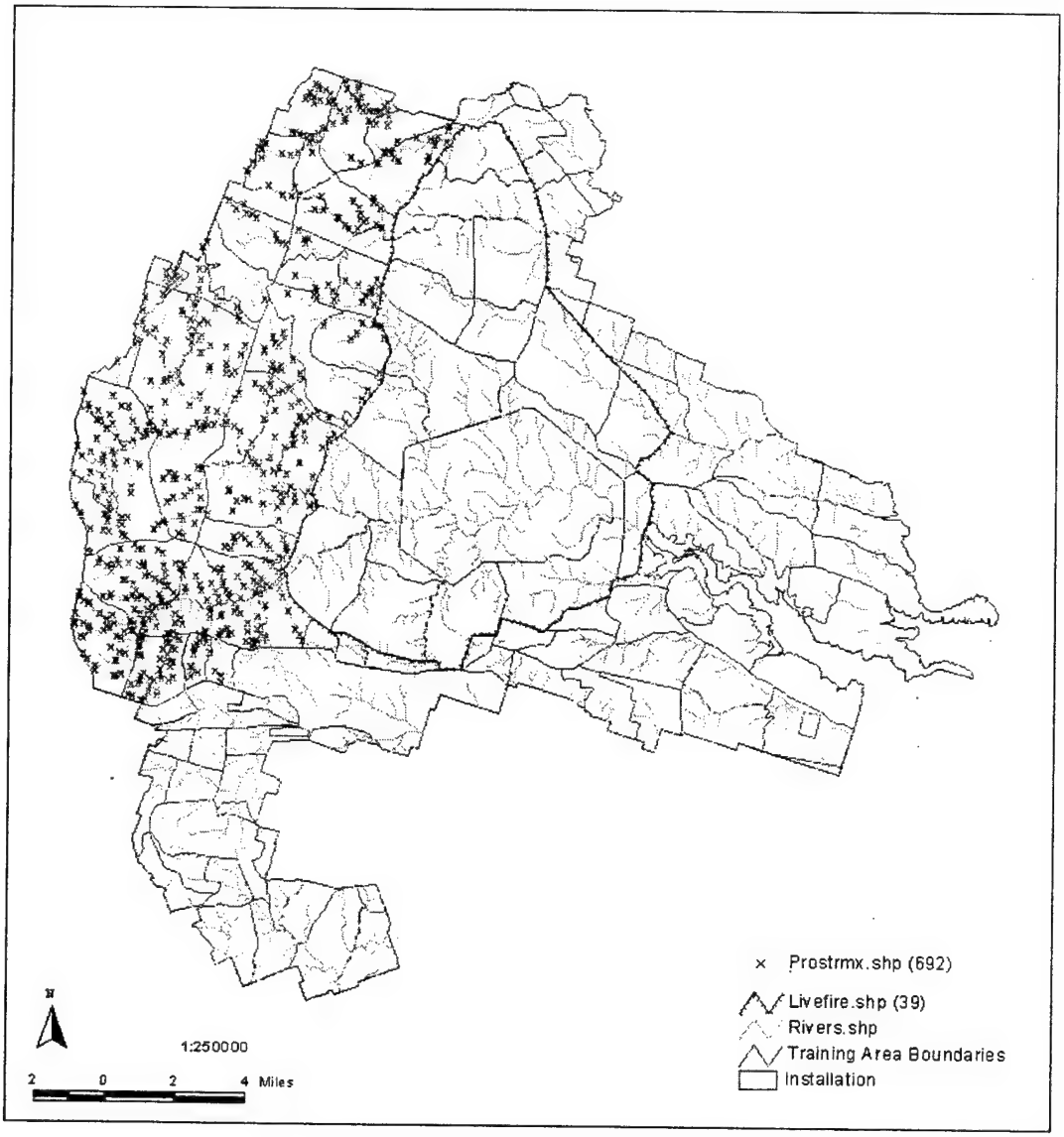
Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Fort Hood Stream Crossings - Prostrmx Only

Initial Data Set, April1999

Map Layout #11



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Rural Drop Locations Data (See Map Layout #12)

One relevant file was delivered in a May FTP transfer:

Teledrop.shp 59 point features

The following description accompanied this file:

"digitized from the 1991 NIMA MIM – 1:50,000 scale"

Teledrop contains point features presumably defining the locations of drop lines within the installation. Attributes in the data table indicate the data probably came from an ARC/INFO source:

area, perimeter (all set to 0)
teledrop_ (ranges from 0 to 62)
teledrop_i (ranges from 0 to 62, matches teledrop_)
data (all set to "#n", where n is a value ranging from 0 and 54)

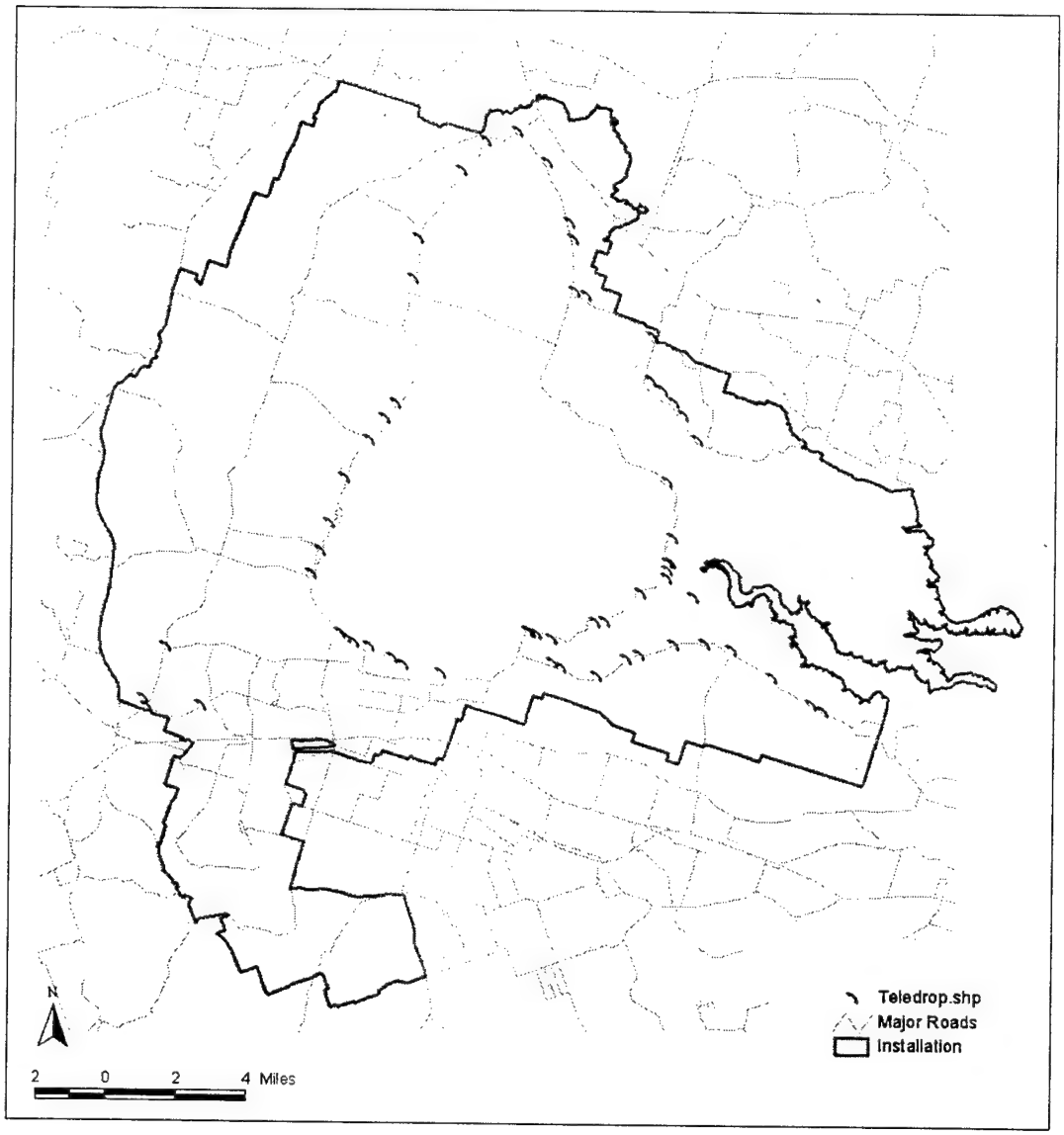
The value for data field is generally not unique. Several points share the same reference; for example, there are four points labeled as "#10". Points with the same data label do not seem to be spatially related.

The majority of the drop line sites are located along the main road encircling the live fire area. Twelve sites are located along North Nolan Road, and four sites are located to the southwest in training areas 33 and 31.

Fort Hood Rural Drop Locations

Initial Data Set, April 1999

Map Layout #12



Prepared for:
Quality Assurance/Quality Control Procedures
for Fort Hood Geospatial Data

Geographic Modeling Systems Laboratory
University of Illinois

Construction Engineering Research Laboratory
US Army Corps of Engineers

Roads

Nine applicable roads data sets in FTP transfer:

Highways.shp 67 polylines
Roadmainold.shp 752 polylines
Roadmajall.shp 2729 polylines
Roads.shp 16944 polylines
Roads_detailed.shp 76,326 polylines
Roads_imp.shp 600 polylines
Roads_LTAM.shp 166 polylines
Roads_unpaved.shp 21869 polylines
Roadssec.shp 140 polylines
Streets.shp 25592 polylines

Highways contains major roads proximate to Fort Hood. Attributes in the data table indicate that the data probably came from a GRASS source:

Seq_num
Grass_line
Cat (contains values of mclennan count, falls count, milam count, williamson count, travis count, burnet count, lampasas count, hamilton count, coryell count, bell count, bell count, citie, highway, no data or a blank field)

There was no description of this data set in the documentation. The extents of this data set are much larger than the boundary of the installation. Of the 12 or so highways in this data set, only 4 of them are actually within or immediately adjacent to the installation.

Roadmainold contains roads within the Fort Hood boundary. Attributes in the data table indicate the data probably came from an ARC/INFO source:

Roadmainold (ranges from 1 to 752)
fnode_ (ranges from 1 to 760)
tnode_ (ranges from 2 to 759
lpoly_, rpoly (all set to 0)
lenght (ranges from 12.277 to 3330.770)
roadspatch (ranges from 1 to 752)
roadspatch (ranges from 1 to 754)

According to the documentation, this data set contains "major roads on Fort Hood (old layer), clipped to boundary". *Roadsmainold* contains the features from *highways* that fall within the boundary of the installation, but the locations and geometries are different.

Roadmajall contains a selection of roads that extend outside the Fort Hood boundary. Attributes in the table indicate that data probably came from an ARC/INFO source:

fnode_ (ranges from 6 to 2610)
tnode_ (ranges from 3 to 2599
lpoly_, rpoly (all set to 0)
length (ranges from 13.789 to 6196.132
roads1_3al (range from 1 to 2729)
roads1_3al (range from 1 to 2754)

According to the documentation, this data set contains "major roads on and in vicinity of Fort Hood (old layer)".

Roadmajall appears to contain the same roads as r*oadmainold*, except that r*oadmainold* was clipped with the Fort Hood installation boundary file (an older version of the boundary, and there are some small errors in length of clipped segments). A spatial join between r*oadmainold* and r*oadmajall* found 702 matches. The unmatched lines occur at the installation boundary.

Roads contains roads that extend outside the Fort Hood boundary. The attributes in the data table indicate that data came from a DLG source:

ID (ranges from 4 to 6259)
Major1 (contain values 170, 172, 173, 174, 179)
Minor1 (contain values 9, 35, 36, 53, 84, 93, 107, 116, 121, 183, 184, 185, 190, 201, 203, 205, 209, 210, 236, 317, 402, 436, 439, 440, 518, 580, 602, 605, 607, 817, 929, 930, 931, 938, 1113, 1123, 1237, 1241, 1741, 1783, 1829, 1996, 2305, 2409, 2410, 2412, 2416, 2484, 2657, 2671, 3046, 3047)
Major2 (contain values 0, 170, 172, 173, 174, 179)
Minor2 (contain values 0, 9, 35, 36, 81, 84, 93, 107, 116, 185,190, 201. 203, 205, 210, 211, 215, 253, 317, 440, 605, 607, 1237, 1670, 1783, 1996, 2271, 2410, 2416, 2483, 2484, 2601, 2671, 2808, 3170, 3219)
Major3, minor3 (all set to 0)
Major4, minor4 (all set to 0)
Major5, minor5 (all set to 0)
Major6, minor (all set to 0)

There was no description of this data set in the documentation. This data set is different in that it defines roads not by their centerline but with two linear features, presumably representing road width.

A spatial join between *roadmajall* and *roads* found no matches, although a visual inspection shows that the same features from *roadmajall* are represented in *roads*. A detailed manual inspection indicates that the spatial join failed because the features are shifted by approximately 1 meter, probably caused by different projection conversion processes.

Roads_detailed contains polylines for some major roads throughout the installation and many minor roads in and around the cantonment areas. Attributes in the data table indicate the data came from a CAD source:

Entity (contains values of "arc", "line" or "line string")
Layer (text field, contains values of 3, 7, 9, 13, 15, 16, 19, 23, 24, 28, 31, 39, 52, 53, 55, 56, 61, 63)
Level (same as layer but numeric field)
Elevation (all set to 0.00000)
Color (contains values set to 0, 1, 3, 4, 6, 106, and 155)
Mslink_dmr (contains values of 0, 201262, 201266, 202043, and 202112
Mslink_dmr (all fields are blank)

There was no description of this data set in the documentation. The set of roads represented in this data set is very extensive. It seems to include any paved vehicle transportation route, including cul de sacs, parking lots, and driveways. This data set is also different in that it defines roads not by their centerline but with two linear boundaries, presumably representing road width.

Roads_imp contains polylines for a subset of roads within Fort Hood boundary. Attributes in the data table indicate the data came from an Arc/Info source:

Fnode_ (values range from 0 to 715, with many duplicate 0 values)
Tnode_ (values range from 0 to 717, with many duplicate 0 values)
Lpoly_ (all set to 0)
Rpoly_ (all set to 0)
Length (values range from 12.277 to 3294.481)
Roadspatch (values range from 0 to 711, with many duplicate 0 values)
Roadspatch (values range from 0 to 753, with many duplicate 0 values)

According to the documentation, this data set contains "improved roads on Fort Hood (improved paved surface)". A spatial join between *roads_imp* and *roadmajall* found 527 matches. Some of the unmatched features are not represented in *roadmajall*, while others appear to be different geometries.

Roads_LRAM contains polyline features defining roads on the western side of Fort Hood. Attributes in the data table indicate the data came from an Arc/Info source:

fnode_ (values range from 1 to 165)
tnode_ (values range from 2 to 171)
lpoly_, rpoly (all set to 0)
length (values range from 4.481 to 5311.599)
roadpro_ (values range from 1 to 166)
roadpro_id (values range from 1 to 156)

According to the documentation, this data set contains proposed and completed improved LRAM tank trails.

Roads_unpaved contains polyline features defining roads, throughout the installation, with heavy concentration of roads in the cantonment areas. Attributes in the data table indicate the data came from a CAD source:

Mslink (values include 0, 201282, 201286, and 203548)
Name (values include other; parking, unpaved; road, unpaved; and shoulder, unpaved)

There was no description of this data set in the documentation. It appears to be closely related to *roads_detailed*, even though it contains none of the attributes. The features are often, but not always, represented as areas rather than centerlines. Second, a detailed visual inspection showed that the features coincide across the two data sets. There are many instances where features meet at intersections, and there appears to be some duplication. However, the feature locations are slightly offset, indicating that the data sets underwent different transformation processes.

Roadssec contains polyline features defining a limited set of short road segments within Fort Hood boundary. Attributes in the data table indicate the data came from an Arc/Info source:

Fnode_ (range from 0 to 147, some duplicate numbers)
Tnode_ (range from 1 to 146, some duplicate numbers)

Lpoly_ (all set to 0)
Rpoly_ (all set to 0)
Length (range from 25,023 to 11063.999)
Roadssec_ (range from 0 to 138, 0 appears twice)
Roadssec_I (range from 0 to 136, 0, 123, and 136 appear twice)

According to the documentation, this data set contains "secondary roads within Fort Hood boundary".

A spatial join between *roadssec* and *roadmajall* found 122 matches. Some of the unmatched features appear to be caused by differences in segment lengths due to clipping; one main segment appears to be a different definition of the road geometry.

Roadssec and *roads_imp* appear to be subsets of the same parent. A spatial join between the two found 25 matches, mostly small segments located at points where larger road segments from each data set meet. In addition, the main road at the southeast corner of the installation seems to be represented in both files but with different geometries.

Although the geometries are very different, there appears to be overlap in terms of the features being represented in *roads_unpaved*, *roads_imp,* and *roadssec*.

CERL Distribution

Chief of Engineers
ATTN: CEHEC-IM-LH (2)
ATTN: HECSA Mailroom (2)
ATTN: CECC-R

Assistant Chief of Staff for Installation Management
ATTN: ODEP 20310-0600

Commander, Army Training Support Center
ATTN: ATIC-ATML-LM 23604

Engineer Research and Development Center (Libraries)
ATTN: ERDC, Vicksburg, MS
ATTN: Cold Regions Research, Hanover, NH
ATTN: Topographic Engineering Center, Alexandria, VA

Defense Tech Info Center 22304
ATTN: DTIC-O

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)
08-2000

2. REPORT TYPE
Final

3. DATES COVERED (From - To)

4. TITLE AND SUBTITLE
Quality Assurance/Quality Control Procedures for ITAM GIS Databases

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Douglas M. Johnston, Diane M. Timlin, Diane L. Szafoni, Jason J. Casanova, and Kelly M. Dilks

5d. PROJECT NUMBER
62720A917

5e. TASK NUMBER

5f. WORK UNIT NUMBER
BF8

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
P.O. Box 9005
Champaign, IL 61826-9005

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDC/CERL TR-00-20

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Assistant Chief of Staff for Installation Management
7701 Telegraph Road, Casey Bldg.
Alexandria, VA 22310-3862

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Copies are available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.

14. ABSTRACT

Geographic information creates unique challenges and opportunities for realizing the mission objectives of the U.S. Army, both in tactical operations and in readiness preparation. It is estimated the Federal Government spends over $3 billion on spatial data in each fiscal year and Geographic Information Systems (GISs) have been implemented at nearly every U.S. Army installation in the United States. In spite of the large investment in digital geographic information and systems, numerous challenges to effective creation, analysis, and delivery of geographic information exist.

This research used a combination of theoretical work on the nature of geospatial data quality and a case study to assess the requirements for conducting a data quality assessment. The project develops and tests procedures for performing assessment of the quality of an existing data set (specifically, selected data from the Integrated Training Area Management (ITAM) GIS database in use at Fort Hood, TX). The procedures are intended to be generalizable to other installation ITAM data sets.

15. SUBJECT TERMS
Fort Hood, TX; Integrated Training Area Management (ITAM); Geographic Information Systems (GIS); land management; military training; quality assurance/quality control

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kelly M. Dilks
Unclassified	Unclassified	Unclassified	SAR	132	19b. TELEPHONE NUMBER (include area code) (217)352-6511 x7472